UNIVERSITY OF LIVERPOOL

Antimalarial Drug Design: Targeting the *Plasmodium falciparum* Cytochrome bc_1 Complex through Computational Modelling, Chemical Synthesis and Biological Testing

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Alexandre Simon Lawrenson

June 2012

Declaration

This thesis is the result of my own work. Research was carried out in the Department of Chemistry and the School of Tropical Medicine at The University of Liverpool. The material contained in this thesis has not been presented, nor is currently being presented, either wholly or in part for any other degree qualification.

Acknowledgements

When asked, I tell people a PhD is an odd thing. At times it feels like the loneliest thing you can do, battling an onslaught of literature and learning in the hopes of making a slight contribution towards your field. However, after four years I can say that a PhD is much more than the sum of its parts, and I've matured more during the last four years than any other period in my life, thanks in large part to the support system which has been around me all along.

For that reason this thesis does not just belong to me, but to a number of people who have contributed in many different ways. Academically I must begin by thanking the two people I've been in prolonged contact with the most, Emma Yang and Raman Sharma. Thank you for the day to day banter and support in the office, particularly Raman for your selfless sharing of your experiences and knowledge. You've both made it a joy to come into work, and I'll always see you as friends, not colleagues. The PON group for your support and inclusion, and Professor Paul O'Neill for your teachings over the years, and the opportunity to work on such exciting projects. To Zeyn Hyder for your encouragement in the lab, as well as keeping me amused when I was bumbling my way through a column. To Ally Shone for your continued patience and support during my time over at LSTM, as well as all the other guys over there for helping out a chemist who will always think in kilograms, not nanomoles. Additionally, I would like to thank all the technical staff at The University of Liverpool for their analytical services and help, as well as to express my gratitude to the EPSRC for funding these studies.

For more personal reasons I must thank several people who have made this not only a time of learning, but also a time of fun. To Andrea Laybourn, Cate Cropper, Gemma De Lanerolle and Paul Wiper. Be it our incessant tea and dinner breaks which have helped us through the long days, or simply belly laughing at the numerous embarrassing situations we have all found ourselves in, it's been a pleasure. Huge thanks to Amy Taylor for your much needed comic relief and inspiration, and also to my two closest friends, Carly Brooke and Laura Clews. Two wonderful people who have been there through the good and the bad without judgement. We've all grown up together and will continue to do so.

To my brother Josh, a man of few words who is without a doubt the bravest person I know. To my auntie and uncle, David and Aileen. Your frequent dinner invitations during my poorer student days meant more to me that I think you know, and your generosity is something I aspire to, as is your warmth and caring. To mum. For better or worse I've always been myself with you. Your love has been unconditional and total, and as a result I've been able to do so many amazing things. Thank you. To William Roscoe. Your total lack of interest in science makes me love you all the more. Without a clear balance between work and home I don't know how I would have managed. You're not just my partner but my best mate, and I can't wait to share everything with you.

Finally, the biggest thank you goes without a doubt to Dr Neil Berry. No one could ask for a more tolerant, willing and all round superstar supervisor. From day one you made me feel totally at ease and confident that I really could do this. I hope I've been a good student and that you're aware of my appreciation for this incredible opportunity.

Contents

Title Page		i
Declaration		ii
Acknowledgments		iii
Contents		iv
Abstract		v
Publications To Date		vi
Abbreviations		vii
Chapter I	Introduction	1
Chapter II	Ligand Based Virtual Screening Methods	84
Chapter III	Ligand Based Virtual Screening Scoring & Selection	163
Chapter IV	Ligand Based Virtual Screening Testing & Analysis	197
Chapter V	Molecular Docking Studies	229
Chapter VI	Quantitative Structure Activity Relationships	298
Chapter VII	Synthesis of Quinolone Antimalarials	383
Conclusions & Future Work		416
Experimental Chapter		421
Appendix		454

Abstract

Malaria is a life-threatening disease which is responsible for roughly one million deaths annually. Previous successes in attempting to eradicate the disease have only been short lived, owing to the increased development of resistance in the parasite. There is a continued need for novel compounds which act at novel therapeutic targets, with the *Plasmodium falciparum* cytochrome bc_1 complex (*Pf*bc_1) representing one such target. Its inhibition halts the biochemical generation of ATP, thus resulting in parasite cell death. Work described in this thesis was concerned with utilising molecular modelling, synthesis and biological testing to develop novel antimalarial compounds, which selectively inhibit this target.

The structural details of a number of compounds known to be active or inactive against *Pf*bc_1 were used in combination with six different ligand based virtual screening techniques, and applied to the ZINC lead like library of compounds to identify potential chemotypes active against malaria. These methods included fingerprint similarity searching, principal component analysis, and naïve Bayesian classification. The hits from each of these methods were merged and formed part of a consensus analysis in which compounds identified across several methods were deemed of more interest than those which appeared less frequently. Each molecule was given a score based on its number of occurrences in the virtual screening methods and also its physicochemical properties. Compounds were filtered to remove those with unfavourable chemical properties, or which contained known toxicophores. 19 compounds were ultimately purchased and tested *in vitro* against the 3D7 strain of the malaria parasite. 5 of the compounds reported single digit μM IC_{50} values, with each containing novel structural chemotypes. The lead candidate contained a benzothiazole core, and reported an IC_{50} value against 3D7 of 4.53 ± 1.86 μM. Additional testing showed the compounds to be inactive against bovine bc_1, which is promising as strong bovine bc_1 inhibition has been shown to be indicative of cardiotoxicity in humans.

Molecular docking was extensively employed to rationalise the activity of *Pf*bc_1 inhibitors such as atovaquone and HDQ. A number of quinolone containing compounds were also subject to docking, with key observations made with regard to interactions thought to be crucial to their antimalarial activity. The hits from LBVS were also the focus of docking, further supporting their potential as *Pf*bc_1 inhibitors.

QSARs were developed for a series of 4-aminoquinoline compounds which had been tested against both the NF54 and K1 strains of malaria. MLR, PLS and *k*NN machine learning methods were investigated, with molecular descriptors contained within valid models interpreted. Significant models were identified and shown to have strong predictive abilities for both strains. QSAR models were similarly developed for a series of thiazolide compounds with activity against hepatitis C. SVM was found to give a significant model which was able to predict the cell safety of the thiazolide derivatives.

The rational design of the novel pyrroloquinolone chemotype led to the synthesis of 7 synthetic analogues to investigate its SAR, via alkylation and Winterfeldt oxidation reactions. The compounds reported 3D7 activity values between 75 nM and 1.02 μM, with molecular docking supporting their potential for Q_o binding and thus *Pf*bc_1 inhibition.

Publications To Date

Thiazolides as novel antiviral agents. 2. Inhibition of hepatitis C virus replication

A. V. Stachulski, C. Pidathala, E. C. Row, R. Sharma, N. G. Berry, A. S. Lawrenson, S. L. Moores, M. Iqbal, J. Bentley, S. A. Allman, G. Edwards, A. Helm, J. Hellier, B. E. Korba, J. E. Semple and J.-F. Rossignol, *Journal of Medicinal Chemistry*, 2011, **54**, 8670-8680.

Generation of quinolone antimalarials targeting the Plasmodium falciparum mitochondrial respiratory chain for the treatment and prophylaxis of malaria

G. A. Biagini, N. Fisher, A. E. Shone, M. A. Mubaraki, A. Srivastava, A. Hill, T. Antoine, A. J. Warman, J. Davies, C. Pidathala, R. K. Amewu, S. C. Leung, R. Sharma, P. Gibbons, D. W. Hong, B. Pacorel, A. S. Lawrenson, S. Charoensutthivarakul, L. Taylor, O. Berger, A. Mbekeani, P. A. Stocks, G. L. Nixon, J. Chadwick, J. Hemingway, M. J. Delves, R. E. Sinden, A.-M. Zeeman, C. H. M. Kocken, N. G. Berry, P. M. O'Neill and S. A. Ward, *Proceedings of the National Academy of Sciences*, 2012, **109**, 8298-8303.

The development of quinolone esters as novel antimalarial agents targeting the Plasmodium falciparum bc1 protein complex

R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. O'Neill, *MedChemComm*, 2012, **3**, 39-44.

Cytochrome b Mutation Y268S Conferring Atovaquone Resistance Phenotype in Malaria Parasite Results in Reduced Parasite bc1 Catalytic Turnover and Protein Expression

N. Fisher, R. Abd Majid, T. Antoine, M. Al-Helal, A. J. Warman, D. J. Johnson, A. S. Lawrenson, H. Ranson, P. M. O'Neill, S. A. Ward and G. A. Biagini, *The Journal of biological chemistry*, 2012, **287**, 9731-9741.

Identification of Novel Antimalarial Chemotypes via Chemoinformatic Compound Selection Methods for a High-Throughput Screening Program against the Novel Malarial Target, PfNDH2: Increasing Hit Rate via Virtual Screening Methods

R. Sharma, A. S. Lawrenson, N. E. Fisher, A. J. Warman, A. E. Shone, A. Hill, A. Mbekeani, C. Pidathala, R. K. Amewu, S. Leung, P. Gibbons, D. W. Hong, P. Stocks, G. L. Nixon, J. Chadwick, J. Shearer, I. Gowers, D. Cronk, S. P. Parel, P. M. O'Neill, S. A. Ward, G. A. Biagini and N. G. Berry, *Journal of Medicinal Chemistry*, 2012, **55**, 3144-3154.

HDQ, a potent inhibitor of Plasmodium falciparum proliferation, binds to the quinone reduction site of the cytochrome bc_1 complex

C. Vallières, N. Fisher, T. Antoine, M. Al-Helal, P. Stocks, N. G. Berry, A. S. Lawrenson, S. A. Ward, P. M. O'Neill, G. A. Biagini and B. Meunier, *Antimicrobial Agents and Chemotherapy*, 2012, **56**, 3739-3747.

Abbreviations

µM	Micromolar
0D	Zero-dimensional
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
A	Number of atoms (hydrogen excluded)
ACT	Artemisinin combination therapy
ADME	Absorption, distribution, metabolism and excretion
AGNES	Agglomerative nesting
AIBN	2,2'-azobisisobutyronitrile
Ala	Alanine
a_n	PLS corresponding coefficient
Asp	Aspartic acid
ATOV	Atovaquone
ATP	Adenosine triphosphate
BEI	Binding efficiency index
BMI	Body mass index
C	Concentration
c	Constant/ Dissimilarity level
C. pneumoniae	*Chlamydia pneumoniae*
CADD	Computer aided drug design
CADEX	Computer Adjunct Data Evaluator X
calcd	Calculated
CAS	Chemical Abstracts Service
CC_{50}	Concentration at which 50% cell cytotoxicity is observed
CDC	Centres for Disease Control and Prevention
CFDs	Chemical function descriptors
CLARA	Clustering large applications
CoQ	Coenzyme Q_{10}/Ubiquinone
CQ	Chloroquine
CQR	Chloroquine resistant

CQS	Chloroquine sensitive
D6	Chloroquine sensitive and mefloquine resistant *Plasmodium* strain
Da	Daltons
DHFR	Dihydrofolate reductase
DHPS	Dihydropteroate synthase
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DMT	Drug metabolite transporter
DNA	Deoxyribonucleic acid
D_T	*k*NN threshold
DV	Digestive vacuole
E	Equally active
EA	Evolutionary algorithms
ECFP	Extended connectivity fingerprints
E_{clash}	Protein-ligand clash penalty term
E_{cov}	Covalent energy term
EDTA	Ethylenediaminetetraacetic acid
E_{int}	Internal ligand energy torsion term
EPR	Electron paramagnetic resonance
Eq.	Equation
ESS	Explained sum of squares
ETC	Electron transport chain
EtOAc	Ethyl acetate
F	Fisher statistic
FAD	Flavin adenine dinucleotide
$FADH_2$	Reduced flavin adenine dinucleotide
FCFP	Functional class fingerprints
Fig.	Figure
FMN	Flavin mononucleotide
$FMNH_2$	Reduced flavin mononucleotide
FPIX	Ferriprotoporphyrin IX
g	Grams
GA	Genetic algorithm

GA-MLR	Genetic algorithm multiple linear regression
GAs	Genetic algorithms
GDP	Gross domestic product
Glu	Glutamic acid
Gly	Glycine
GOLD	Genetic Optimisation for Ligand Docking
GPCRs	G-protein coupled receptors
GSK	GlaxoSmithKline
HARD	Hardness
HAT	Human African trypanosomiasis
H-bond	Hydrogen bond
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HDQ	1-hydroxy-2-dodecyl-4(1H)-quinolone
His	Histidine
HIV	Human immunodeficiency virus
HQNO	2-heptyl-4-hydroxyquinoline N-oxide
hr	Hour
H_{rot}	Loss of conformational entropy in the ligand upon binding to the protein
HSAB	hard and soft acids and base
HTS	High throughput screening
Hy	Hydrophilic factor
Hz	Hertz
IC_{50}	Concentration required for 50% inhibition of biological process
ID3	Iterative dichotomiser 3
IFN-α	Interferon α
Ile	Isoleucine
ISP	Rieske iron-sulphur protein
JBC	The Journal of Biological Chemistry
JGI5	Mean topological charge index of order5
K	Number of descriptors
KCN	Potassium cyanide

k_i	Coefficient associated with a particular independent variable
*k*NN	*k*-Nearest Neighbour
L	Less active
LBVS	Ligand based virtual screening
LE	Ligand efficiency
Leu	Leucine
LMO	Leave-many-out
log *BB*	Index of blood-brain barrier permeability
log *P*	Partition coefficient between 1-octanol and water
Log S	Log units of molar solubility
LOO	Leave-one-out
LSTM	Liverpool School of Tropical Medicine
Lys	Lysine
m	Gradient
M	More active
m	Multiplet
MDL	Molecular design limited
Met	Methionine
mg	Milligram
mL	Millilitre
MLR	Multiple linear regression
mM/mmol	Millimolar
MMP	Mitochondrial membrane potential
MOPAC	Molecular Orbital PACkage
Mor31e	3D-MoRSE - signal 31 / weighted by atomic Sanderson electronegativities
mtETC	Mitochondrial electron transport chain
MW	Molecular weight
N	Number of data points
n	Number of observations/molecules
N	Number of non-hydrogen atoms
N%	Percentage of N atoms
NAD^+	Nicotinamide adenine dinucleotide

NADH	Reduced nicotinamide adenine dinucleotide
NBS	*N*-bromosuccinimide
N_C	Number of carbon atoms
NCI	National Cancer Institute
nF	Number of F atoms
NF54	Chloroquine sensitive and mefloquine resistant *Plasmodium* strain
nH	Number of hydrogen atoms
N_{Hy}	Number of hydrophilic groups
nM	Nanomolar concentration
nm	Nanometre
NMR	Nuclear magnetic resonance
NQNO	2-*n*-nonyl-4-hydroxyquinoline N-oxide
NTZ	Nitazoxanide
nvar	Number of variables
nX	Number of halogens
p	Number of variables
P. berghei	*Plasmodium berghei*
P. falciparum	*Plasmodium falciparum*
P. knowlesi	*Plasmodium knowlesi*
P. malariae	*Plasmodium malariae*
P. ovale	*Plasmodium ovale*
P. vivax	*Plasmodium vivax*
PCA	Principal component analysis
PCR	Principal component regression
PCs	Principle components
PDB	Protein data bank
PDE	Phosphodiesterase
Pfbc_1	*Plasmodium falciparum* cytochrome bc_1 complex
Pfcrt	*Plasmodium falciparum* chloroquine resistance transporter
*Pf*DHFR	*Plasmodium falciparum* dihydrofolate reductase
*Pf*DHODH	*Plasmodium falciparum* dihydroorotate dehydrogenase
*Pf*NDH2	*Plasmodium falciparum* type II NADH dehydrogenase
PFT	Protein farnesyltransferase

Phe	Phenylalanine
PLS	Partial least squares
pM	Picomolar concentration
PMII	Plasmepsin II
ppm	Parts per million
PRESS	Predictive residual sum of squares
Pro	Proline
PSA	Polar surface area
PTR1	Pteridine reductase 1
q^2	Cross-validated r^2
q^2_{ext}	External cross-validated r^2
Q	Ubiquinone
QH2	Ubiquinol
Q_i	Quinone reduction site
Q_o	Quinol oxidation site
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationship
r^2	Squared correlation coefficient
r_0^2	Correlation coefficient of regression line through the origin
r^2_{adj}	Adjusted squared correlation coefficient
R	Correlation coefficient
RBC	Red blood cell
RBF	Radial basis function
RBs	Rotatable bonds
RMSD	Root mean square deviation
RSS	Residual sum of squares
RT	Room temperature
s	Standard error of prediction
s	Singlet
S. cerevisiae	*Saccharomyces cerevisiae*
SAHN	Sequential agglomerative hierarchical non-overlapping
SAR	Structure activity relationship
SBVS	Structure based virtual screening

SD	Standard deviation
SDH	Succinate dehydrogenase
Ser	Serine
S_{hb_ext}	Protein-ligand hydrogen bond energy (external H-bond)
S_{hbond}	Score for hydrogen-bonding interactions
S_{lipo}	Score for lipophilic interactions
S_{metal}	Score for acceptor-metal interactions
SMILES	Simplified Molecular Input Line Entry System
SMOTE	Synthetic minority oversampling technique
SOC	Standard of care
S_{tor_int}	Ligand torsional strain energy (internal torsion)
S_{vdw_ext}	Protein-ligand van der Waals energy (external vdw)
S_{vdw_int}	Intramolecular strain in the ligand vdw energy (internal vdw)
SVM	Support vector machines
t	Threshold dissimilarity/Triplet
THF	Tetrahydrofuran
THPI	1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole
t_i	Latent variables
TLC	Thin layer chromatography
Trp	Tryptophan
TSS	Total sum of squares
Tyr	Tyrosine
u	Unknown object
UFS	Unsupervised forward selection
V	Total volume occupied
v/v	Volume to Volume
vdw	Van der Waals
VSS	Virtual Screening Score
WEKA	Waikato environment for knowledge analysis
WHO	World Health Organisation
x	Independent variable
y	Dependent variable
y_i	Observed response for dependent variable

$\hat{y}_i$	Predicted response for dependent variable
$\bar{y}$	Average values of observed dependent variable/Average Euclidean distance between compound
$\hat{\bar{y}}$	Average values of predicted dependent variable
$\bar{y}_{tr}$	Average response across the entire training set for dependent variable
$y_{calc,i}$	Calculated value of the dependant variable from regression equation
$y_{pred,i}$	Predicted value of the dependant variable from regression equation
Z	Arbitrary parameter to control the significance level (*k*NN)
$\langle x \rangle$	Mean of the independent variables
$\langle y \rangle$	Mean of the dependent variables
ΔG	Gibbs free energy/Coefficients derived from MLR analysis
$\Delta G_{binding}$	Total free energy of binding change
π	Hydrophobicity constant
σ	Hammett substitution parameter/Standard deviation of Euclidean distance

Chapter I

Introduction

1. Introduction 4

1.1 Malaria 4

1.2 *Plasmodium* Life Cycle 6

1.3 Antimalarial Chemotherapy 9

1.3.1 Chloroquine (4-Aminoquinolines) 11

1.3.1.1 Chloroquine Resistance 12

1.3.2 Artemisinins and Endoperoxides 14

1.3.2.1 Artemisinin Resistance 16

1.3.3 Chimeric Compounds 18

1.3.4 Antifolates 20

1.3.5 Mitochondrial Electron Transport Chain Inhibitors 21

1.3.5.1 Complex III - Cytochrome bc_1 Complex 25

1.3.5.2 Atovaquone 28

1.3.5.2.1 Atovaquone Resistance 30

1.4 Drug Design and Discovery 31

1.5 Molecular Design Loop 34

1.6 Computational Chemistry 35

1.7 Chemoinformatics 36

1.8 Virtual Screening 37

1.8.1 Ligand Based Virtual Screening 38

1.8.1.1 Similarity Searching 39

1.8.1.1.1 Tanimoto Coefficient 40

1.8.1.2 Pharmacophore Mapping 41

1.8.1.3 Quantitative Structure Activity Relationship 43

1.8.1.3.1 Molecular Descriptors 44

1.8.1.3.2 Multiple Linear Regression 46

1.8.1.4 Machine Learning Methods 49

1.8.1.5 Ligand Based Virtual Screening Successes 51

1.8.2 Structure Based Virtual Screening 52

1.8.2.1 Molecular Docking 53

1.8.2.1.1 Scoring Functions 57

1.8.2.2 Structure Based Virtual Screening Successes 60

1.9 Chemical Synthesis 62

1.9.1 Synthetic Discovery of Chloroquine 62

1.9.1.1 Chloroquine Analogues 64

1.9.2 Synthesis of Novel Antimalarial Compounds 66

1.9.2.1 Hydroxynaphthoquinones 66

1.9.2.2 Pyridones 67

1.9.2.3 Quinolones 68

1.10 Biological Testing 69

1.10.1 Bioassays 69

1.11 Aims of this Thesis 71

1.12 References 73

1. Introduction

1.1 Malaria

Malaria is a life-threatening disease which is responsible for roughly one million deaths and some half a billion clinical episodes each year.[1-3] In Africa, a child dies every 45 seconds of the disease. It is transmitted through the bite of the female *anopheles* mosquito which acts as a vector for the malaria parasite. Approximately 40% of the world's population are exposed to the risk of malaria, mainly those in tropical and sub-tropical countries.[4, 5] This is due to the significant amount of rainfall and consistently high temperatures in these regions. These features combined with areas of stagnant waters in which the larvae can mature provides optimum conditions for continuous breeding of the mosquitoes, and thus persistent transmission of the disease.[6] The mosquitoes only feed during the night (dusk till dawn) when bite prevention is imperative.[7] Previous successes in attempting to eradicate the disease were only short lived due to increasing resistance of the mosquito to insecticides,[8] and of the parasite to established drugs.[9]

Malaria is caused by a parasite called *Plasmodium*, of which there are four main human species: *Plasmodium falciparum* (*P. falciparum*), *Plasmodium vivax* (*P. vivax*), *Plasmodium ovale* (*P. ovale*), and *Plasmodium malariae* (*P. malariae*).[10] Recently there have been reported cases of human infection with a fifth species, *Plasmodium knowlesi* (*P. knowlesi*), which is usually found in monkeys.[11, 12] *P. vivax* and *P. falciparum* are the most common of the *Plasmodium* species, but *P. falciparum* is by far the deadliest, with higher mortality rates than the other species.[13, 14]

Malaria is an acute febrile illness, with symptoms of the disease only becoming apparent around ten to fifteen days after the infective mosquito bite. Initial symptoms include fever, chills, headache and vomiting, and thus may be mild and unrecognisable as malaria. However, if not treated, *P. falciparum* quickly progresses to severe illness and may lead to death. Children with severe disease often develop other conditions such as severe anaemia, respiratory distress, or cerebral malaria, and in adults multi-organ involvement is frequent, with the blood supply to vital organs becoming disrupted.[1]

Those living in endemic areas progressively acquire a semi-immune status following repeated bites, reducing acute infection symptoms and disease severity.[15] However, immunity may be partially lost after years in non-endemic countries, as occurs frequently in migrants from endemic countries that move to Europe or North America.[16] Therefore, those individuals have an elevated risk of illness when they return to their country of original to visit friends and family.[17]

Malaria is geographically specific due to the ecological conditions required for the mosquito vector, yet countries that have eliminated its presence generally experienced economic growth in the subsequent years compared to those of neighbouring countries. The disease poses a huge burden where it is endemic, and can cut gross domestic product (GDP) by as much as 1.3%.[1, 18] These aggregated annual losses have resulted in substantial differences in GDP between countries with and without malaria, and in some heavily burdened countries, the disease accounts for up to 40% of all public health expenditures.

The disease also poses worldwide implications due to increases in air travel, with those from malaria free areas of the world especially vulnerable when they become

infected.[1] About 30,000 international travellers develop malaria each year, resulting in around 150 deaths.[19] It is vital that travellers to high risk areas be educated about personal preventative measures that can be used to reduce the possibility of infection, such as wearing long sleeve shirts, sleeping in air conditioned rooms, using mosquito repellents and insecticides, as well as using bed nets and window screens.[20]

1.2 *Plasmodium* Life Cycle

The life cycle of the malaria parasite consists of a sexual cycle which takes place within the mosquito, and an asexual cycle which occurs in man, as depicted by figure 1.1.[10, 21-23]

This text box is where the unabridged thesis included the following third party copyrighted material:

(http://www.cdc.gov/malaria/about/biology/).

Fig. 1.1 *Plasmodium* life cycle. Retrieved from the Centre for Disease Control and Prevention (CDC). (Available at http://www.cdc.gov/malaria/about/biology/)

Within the asexual cycle there is an exo-erythrocytic and an erythrocytic phase. The exo-erythrocytic stage involves infection of the liver, whereas the erythrocytic phase

is concerned with infection of the red blood cells (RBCs). Following the blood meal of an infected female mosquito, *Plasmodium* sporozoites from the mosquitoes saliva enter the bloodstream, where within thirty minutes they move to and infect the hepatocytes (liver cells). Over the next ten to fourteen days they undergo an exo-erythrocytic stage of development and multiplication into liver schizonts. These schizonts then rupture releasing merozoites into the bloodstream, which can then go on to invade the RBCs.

When the merozoites enter the RBCs the erythrocytic phase of the life cycle begins. This is again an asexual stage of multiplication, where the merozoites develop into motile intracellular ring forms of the parasite termed trophozoites. During maturation within the RBC, the parasite remodels the host cell by inserting parasite proteins and phospholipids into the red blood cell membrane. The host's haemoglobin is digested and transported to the parasite's food vacuole where it provides a source of amino acids. Free haem is a by-product of this reaction and is ordinarily toxic to the *Plasmodium*; however, it is rendered harmless by polymerisation to *haemozoin*,[24] a non-toxic molecule which collects as an insoluble crystal in the parasite food vacuole.

Following mitotic replication of its nucleus, the trophozoites mature into schizonts, whose rapid growth and division releases additional merozoites into the bloodstream when the schizont ruptures. It is this rupturing of the schizont and release of further merozoites into the blood which gives rise to the classic symptom of malaria, which is the cyclical occurrence of sudden coldness followed by rigor and then spiking fever and sweating, lasting anywhere between four to six hours. These merozoites are released into the bloodstream and can bind to and enter further RBCs, repeating the erythrocytic cycle.

This cycle occurs in simultaneous waves (from invading merozoites to rupturing schizont) every two days in *P. vivax*, *P. ovale* and *P. falciparum*, and every three days for *P. malariae*. For *P. knowlesi* this process repeats roughly every day. However, in non-immune patients, particularly those infected with *P. falciparum*, parasites tend to mature asynchronously resulting in irregular fever patterns.[25]

On entering the liver cells, some sporozoites in *P. vivax* and *P. ovale* infections may not develop into exo-erythrocytic phase merozoites, but instead become dormant liver parasites termed hypnozoites. [26, 27] Hypnozoites are responsible for long incubation periods and late relapses in the disease. After a period of dormancy, the hypnozoites reactivate and release merozoites weeks to months after the initial prophylaxis and treatment of the primary infection, reactivating the exo-erythrocytic cycle. Hypnozoites pose a threat when it comes to eradicating the disease due to the potential relapses in malaria.[28]

During the erythrocytic phase, some merozoites can differentiate into gametocytes which are a sexual form of the *Plasmodium*. These male and female forms of the parasite can only complete the sex cycle if they are taken from the blood of an infected person when the mosquito takes its blood meal. Since the gametocytes are formed in the blood of the vertebrate host, this host is in fact the definitive host of the disease. The parasites multiplication in the mosquito is known as the sporogonic phase. Over a period of two weeks the sexual cycle is completed through fertilisation of the female gametocyte by the male gametocyte, forming a zygote. These zygotes become motile and elongated and are termed ookinetes, at which point they can invade the mid-gut wall of the mosquito and develop into oocysts. The oocysts can then grow, rupture, and release sporozoites which migrate to the

mosquitoes salivary glands, ready to infect human hosts and continue the malaria life cycle through subsequent bites.

1.3 Antimalarial Chemotherapy

An early diagnosis of malaria and onset of treatment reduces the disease and prevents death, whilst also contributing to a reduction in transmission rates. Malaria is typically diagnosed through the microscopic examination of blood,[29] but both saliva and urine have been investigate as less invasive alternative.[30] However, many areas cannot afford laboratory diagnostic testing so quite often a diagnosis is made at home based on a history of recurrent fevers.[31] This can lead to misdiagnosis and therefore improper treatment for potential underlying conditions such as pneumonia and meningitis. Significant developments in commercial antigen detection methods for malaria diagnosis have been made and field tested in areas where microscopy is unavailable.[32] These tests require only a single drop of blood and are completed within twenty minutes, the results of which are easily interpretable through the presence or absence of a coloured stripe on the testing dipstick. Concerns with this method though are that the results are qualitative not quantitative, and are unable to determine the number of parasites present in the blood, and thus disease progression.

Once a malaria diagnosis has been made, drug chemotherapy has proven to be an effective strategy to treat and cure the disease. Different drugs act on different stages of the *Plasmodium* life cycle, with many being used not only to treat infected individuals, but also in a chemoprophylaxis approach to prevent the disease in travellers.[15, 33] With regard to both treatment and prophylaxis there are a number of options available.

Traditionally, antimalarial drugs were classified based on which stage of the *Plasmodium* life cycle they targeted.[34] Blood schizonticides act on the asexual intraerythrocytic stages of the parasite, and are sufficient to cure *P. falciparum* infections as this *Plasmodium* has no dormant liver phase. Tissue schizonticides kill hepatic schizonts, thus preventing the invasion of erythrocytes. Hypnozoiticides are required though to fully cure *P. vivax* and *P. ovale* infections, as they kill the persistent intrahepatic stage parasites which cause relapses in malaria, and as such should be combined with blood and tissue schizonticides.[35] Finally, gametocytocides destroy intraerythrocytic sexual forms of the parasite, preventing transmission from human to mosquito, thereby halting the *Plasmodium* life cycle.

The first medicine used in the treatment of malaria was found in the powdered bark of the chinchona tree, the active ingredients of which were the quinoline alkaloids quinine and quinidine (fig. 1.2).[34]

Quinine Quinidine

Fig. 1.2 Quinine and quinidine, the first antimalarial agents.

Attempts at synthesising quinine led to the surreptitious discovery of a number of quinoline analogues active against malaria, which in turn drove a number of structure activity relationship (SAR) studies to design more activity antimalarial agents. The end result of these studies was the discovery of chloroquine (CQ), which went on to become the foundation of malaria chemotherapy for at least two decades.[36-39]

1.3.1 Chloroquine (4-Aminoquinolines)

Chloroquine (fig. 1.3) is a 4-aminoquinoline compound and was previously one of the most important synthetic chemotherapeutic agents in history. The World Health Organisation (WHO) once considered it its drug of choice for the Global Eradication Program, and until recently CQ was the cornerstone of antimalarial treatment and prophylaxis.[7, 21]

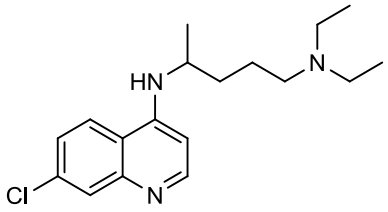


Fig. 1.3 Chloroquine, a 4-aminoquinoline compound.

CQ is a relatively affordable drug that has been widely tested, and which despite having a narrow therapeutic index with a therapeutic dose of only 10 mg kg^{-1}, is reasonably well tolerated owing to its efficacy, tolerability, and safety in pregnancy and childhood.[15, 34] However, long term prophylaxis may lead to irreversible neuromyopathy, retinopathy and other conditions, but these incidences are rare.[40]

The mechanism of action of CQ is still a matter of much debate, but it is commonly agreed that it is a potent blood schizonticidal drug which acts against the erythrocytic forms of all *Plasmodium*.[41, 42] As previously discussed, the parasite consumes large amounts of haemoglobin within the host cell which is shuttled by vesicles to the food vacuole, also called the digestive vacuole (DV). During the digestion of haemoglobin it is broken down into its component peptides which are thought to be exported from the DV, leaving behind ferriprotoporphyrin IX (FPIX), also known as haem.[43-45] Ordinarily this free haem is toxic to the parasite at high concentrations, but the parasite disposes of this hazardous waste through the formation of an

insoluble polymer called haemozoin. CQ acts similar to other 4-aminoquinlines by forming a complex with this free haem, preventing its polymerisation to haemozoin.[46,47]

CQ is a dibasic compound with pK_a values of 10.2 and 8.1 for the diethylamine and quinoline ring nitrogen's respectively.[48] Unprotonated CQ is membrane permeable and thought to distribute equally across all cell compartments. However, CQ becomes trapped in the acidic DV (estimated pH of 5.2-5.8)[49] as a dication, where it can accumulate by some orders of magnitude in its membrane impermeable form.[50] Most NMR and molecular modelling studies[51] seem to suggest that the quinoline part of CQ binds to the porphyrin ring system of haem with face-to-face π staggering, and that this binding allows for the build-up of non-crystalline haem, ultimately leading to parasite cell death.[52] The clinical use of CQ has diminished of late however, due to the continued development of resistance in the malaria parasites.[34]

1.3.1.1 Chloroquine Resistance

CQ was the most widely used antimalarial drug in the 1940's,[53] but by the late 50's (only twelve years after its initial introduction) the first cases of *P. falciparum* CQ resistance began to emerge in Thailand and Cambodia.[15] Due to its massive use, chloroquine resistant (CQR) *Plasmodium* strains developed in many regions, and have successively spread over almost all the malaria endangered parts of the world.[54, 55] Presently, more than 80% of wild isolates are resistant to CQ.[56]

The mechanism of action of CQ resistance has been the subject of much research. It is believed to be removed from its site of action in the DV due to an enhanced efflux of the compound from the parasite vesicles, as a result of increased expression of the

multidrug resistance transporter P-glycoprotein. This influences the sensitivity of the malaria parasites to a diverse range of drugs.[57-59] The primary cause of CQ resistance is a mutation in the *Pfcrt* gene that codes for a protein called the *P. falciparum* chloroquine resistance transporter (*Pfcrt*). This 10-transmembrane domain transport protein belongs to the drug metabolite transporter (DMT) superfamily located in the membrane of the DV. It is proposed that the role of this protein is to transport amino acids or small peptides from haemoglobin degradation, into the cytoplasm.[60]

CQR strains of the parasite all have a threonine residue in place of a lysine at position 76 of the protein.[61, 62] Ordinarily, the positively charge lysine side chain is thought to prevent access of the dicationic form of CQ to the substrate binding site of the transporter, but the K76T mutation replaces this positively charged side chain with a neutral moiety. This alteration in the membrane protein affects the ion trapping effect which CQ activity is reliant upon, by allowing the CQ dication to access the transporter, which in turn decreases the concentration of CQ in the DV down a steep concentration gradient.[48] This leads to a reduction in the overall concentration of CQ at its site of action, and thus a decrease in the sensitivity of the parasite.

Research into the 4-aminoquinoline class of compounds continues to be an active area, with a number of modifications to the CQ structure able to circumvent the aforementioned resistance mechanism.[42, 63, 64] Three main observations to combat resistance have been made: elongation or shortening of the diaminoalkyl side chain; introduction of lipophilic or aromatic moieties into the side chain; dimerization of two 4-aminoquinolines by a linker of variable nature and length. Amodiaquine (fig.

1.4) is an example of one such enhancement of CQ, in which the lipophilicity of the side chain was increased through introduction of an aromatic structure.

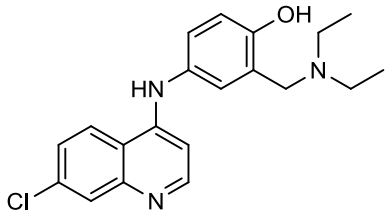


Fig. 1.4 Amodiaquine, a synthetic derivative of CQ.

Though there is a certain degree of cross resistance between amodiaquine and CQ, amodiaquine is effective against low level CQ resistant *P. falciparum*.[40] However, adverse reactions such as hepatotoxicity and agranulocytosis from prolonged prophylaxis have lead to amodiaquine being removed from the market in western countries,[65-68] but the drug may continue to find use in the developing world,[69] and is still an antimalarial recommended by the WHO.[70]

1.3.2 Artemisinins and Endoperoxides

Artemisinins and synthetic peroxides are another class of antimalarial drug whose origins are based in natural product chemistry. Extracts of the herb *Artemisia annua*, also known as sweet wormwood have been used in traditional Chinese medicine for two thousand years for the treatment of fever. The active ingredient is artemisinin (ART, fig. 1.5), a sesquiterpene lactone which was isolated in 1971, and has been used in China for the treatment of malaria since 1972.[71] It is a highly active antimalarial agent with studies reporting IC_{50} values in the nanomolar (nM) concentration activity range.[72, 73] An IC_{50} measurement is simply the concentration of drug required to inhibit 50% of a biological process, giving a quantitative value of a compounds inhibitory activity.

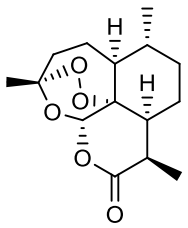


Fig. 1.5 Artemisinin, a sesquiterpene lactone antimalarial.

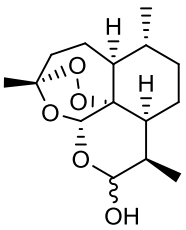
ART has a 1,2,4-trioxane substructure, the endoperoxide of which is believed to be essential for antimalarial activity.[34] The exact mechanism of action of such compounds is still a topic of much debate, but they are believed to act within the parasite DV, where it is proposed that iron(II)-mediated cleavage of the endoperoxide bridge leads to the formation of an oxyl radical, which rapidly rearranges to different C-centred radicals that may be primary or secondary in nature. One or both of these radicals may be the active species, whose formation is believed to be activated by free haem in the DV. The C-centred radicals are thought to react with free haem preventing its detoxification, as well as alkylating to vital parasite proteins, inhibiting a multitude of enzymes and leading to parasite cell death.[45, 74-78]

Artemisinins predominately act against the late ring stages of the *Plasmodium* life cycle in which metabolic activity is highest. However, in contrast to other antimalarials, they also act against the small ring stages present in the erythrocytes a few hours after infection. Additionally, artemisinins are also active against the sexual blood stages of the parasite, reducing rates of parasite transmission in the process.[79] This combined with artemisinins high activity can reduce parasite biomass 10,000 fold in a single asexual cycle, making them the most active and rapidly acting antimalarials known today.[80, 81]

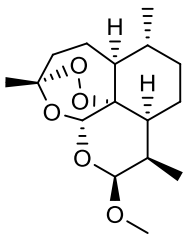
1.3.2.1 Artemisinin Resistance

As was seen with the 4-aminoquinoline class of compounds, there are growing concerns with regard to parasites developing resistance in many parts of the world, and it was thought that the plant derived compound ART remained the only globally effective treatment.[82] However, scientists claim to have found the first evidence of resistance towards ART in parts of Southeast Asia.[4, 83, 84] Emerging resistance is responsible for a recent increase in malaria mortality, particularly in countries which had previously eliminated its presence. In China and Southeast Asia, ART is often used without taking precautions against conditions which may lead to resistance of *Plasmodium* to the drug, so there are concerns that the effectiveness of ART treatment may be reduced in the near future, as seen with the 4-aminoquinolines.[84] Artemisinin combination therapy (ACT) is now the most significant and effective treatment protocol for malaria. However, the combination therapies have shelf lives of as little as three years, which is a major limitation when distributing these drugs.[85]

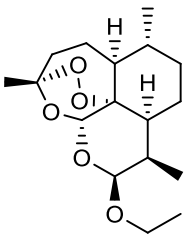
Despite the emergence of resistance to artemisinins it is still very much an active area of research. ART itself is poorly soluble in water and oil so semi-synthetic derivatives have been developed to improve its lipophilicity. Reducing the lactone substructure of ART into a hemiacetal creates dihydroartemisinin (fig. 1.6), which can undergo alkylation to yield artemether and arteether, both characterised by an acetal moiety.



Dihydroartemisinin



Artemether



Arteether

Fig. 1.6 Dihydroartemisinin, artemether and arteether, synthetic peroxides and derivatives of ART.

Artemether and arteether both have antimalarial activity (dihydroartemisinin also exhibits some antimalarial activity). However, artemether is the more prevalent of the two, displaying IC_{50} values in the nM range.[86-88] Artemisinins are often used in combination therapies to improve their half lives and reduce their rate of clearance.[89] An alternative modification of dihydroartemisinin is artesunate (fig. 1.7), in which the hemiacetal OH group can be acylated with succinic acid. Despite the instability of artesunate, the succinic ester group is rapidly cleaved to release the active agent, dihydroartemisinin.

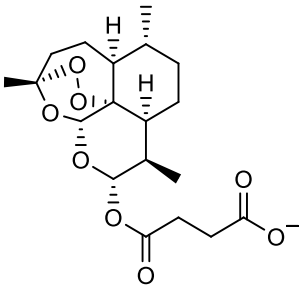


Fig. 1.7 Artesunate, a synthetic peroxide and derivative of ART.

A major limitation with semi-synthetic ART derivatives is that their production relies on a sufficient supply of ART isolated from plants,[34] the extraction of which commonly has a yield of only 0.6%.[90] This, along with the increased use of ACT across the world means that raw material is in short supply.[91]

Once it was clear that the antimalarial activity of ART and its derivates was attributed to the endoperoxide substructure, research began to develop fully synthetic

endoperoxide antimalarials.[78, 92] A complicated synthetic procedure for such endoperoxides often leaves them unsuitable for upscale production, with many often occurring as racemic products with poor pharmacokinetic profiles. However, much work lead to the discovery of OZ-277 (fig. 1.8), also known as arterolane. At first glance its molecular structure may look like an unlikely pharmacological candidate, owing to the presence of an ozonide group and adamantane substituent, but it was the first endoperoxide compound to enter into clinical trials.[93, 94]

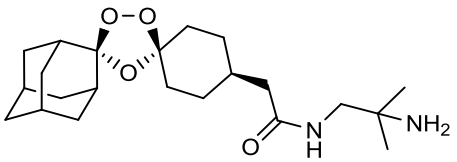


Fig. 1.8 OZ-277/Arterolane, a synthetic endoperoxide antimalarial.

Whilst 4-aminoquinolines and ART derivates and endoperoxides represent some of the most important antimalarial treatment options to date, there are in fact many other options which have been considered.

1.3.3 Chimeric Compounds

With the assumption that 4-aminoquinoline and trioxane compounds act on haem within the *Plasmodium* life cycle, chimeric compounds attempt to incorporate both of these structural moieties into their design.[95, 96] The most promising compound in this trioxaquine series is shown in figure 1.9, and was found to be active in the nM range, alkylating to haem *in vitro*.[97]

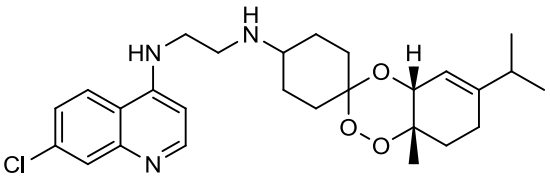


Fig. 1.9 A trioxaquine compound active against *Plasmodium* in the nM range.

Another compound shown to have high nM activity against chloroquine sensitive (CQS) and resistant *Plasmodium* strains[86, 98, 99] was mefloquine (fig. 1.10), a synthetic analogue of quinine (fig. 1.2). However, despite initial clinical successes, resistance has since developed against mefloquine,[88, 100] and though combination therapy is recommended with artesunate (fig. 1.7),[101] pharmacokinetic concerns may lead to the selection of resistant strains.[102, 103] Yet despite concerns with mefloquine, its incorporation into chimeric molecules has lead to some favourable results for this class of compound.

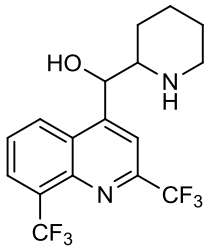


Fig. 1.10 Mefloquine, a synthetic analogue of quinine.

A biologically cleavable chimera of mefloquine and 10-trifluoromethylartemisinin has been prepared (fig. 1.11), showing IC_{50} values in the nM range against a number of parasite strains.[104]

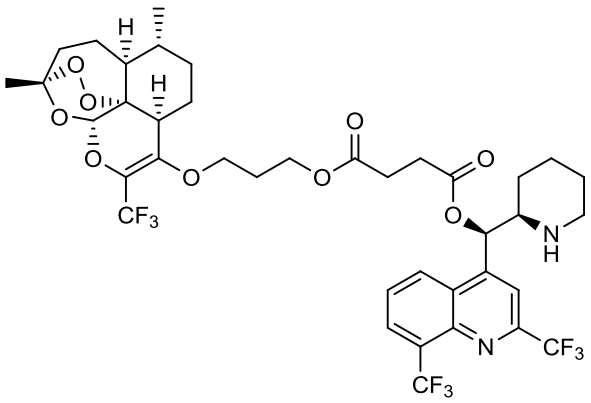


Fig. 1.11 Biologically cleavable chimera of mefloquine and 10-trifluoromethylartemisinin.

1.3.4 Antifolates

Antifolates are another example of antimalarial compounds, with drugs such as proguanil (fig. 1.12) being introduced in the late 1940's for the treatment and prophylaxis of malaria.[34] Tetrahydrofolic acid plays a key role in the biosynthesis of thymine, purine nucleotides, and several amino acids, but humans depend entirely on a dietary intake of dihydrofolic acid, which then undergoes reduction via dihydrofolate reductase (DHFR) to tetrahydrofolic acid. Pathogenic microorganisms however, including *Plasmodia*, synthesise dihydrofolic acid directly from simple precursors, utilising dihydropteroate synthase (DHPS). DHPS is completely absent in humans making it an attractive drug target. Additionally, DHFR is also sufficiently different in *Plasmodia* to that of humans, allowing for the development of selective inhibitors against both enzymes. DHFR and DHPS inhibitors have long been used in the treatment of bacteria and protozoal infections, with their use as antimalarials also well documented.[105, 106]

Proguanil acts as a prodrug, yielding the active metabolite cycloguanil (fig. 1.12) through oxidative ring closure. Cycloguanil is effective against sporozoites acting as a DHFR inhibitor to halt the folate biosynthetic pathway. However, due to widespread use of these compounds resistant strains have begun to emerge.[107, 108] Resistance is thought to have developed through an accumulation of mutations in the *dhfr* gene, which leads to steric clashes when compounds such cycloguanil try to bind with *Pf*DHFR. The folate biosynthetic pathway can therefore continue as normal.

Proguanil Cycloguanil

Fig. 1.12 Prodrug proguanil and its active metabolite cycloguanil.

With emerging resistance across many of the antimalarial classes, and mortality and morbidity rates on the rise, there is a necessary need for continued research into antimalarial drug development. This should include the search for novel drugs which act at novel targets, the results of which may overcome the clinical resistance observed in many marketed antimalarials.[54, 109, 110]

1.3.5 Mitochondrial Electron Transport Chain Inhibitors

In recent years the mitochondrial electron transport chain (mtETC) has been explored for the development of new antimalarials treatments. In higher organisms the electron transport chain (ETC) occurs within the inner mitochondrial membrane, where electron transfer is coupled with the transfer of protons across a membrane. The resulting electrochemical proton gradient is used to generate chemical energy in the form of adenosine triphosphate (ATP), which plays a critical role in respiration.[111, 112]

The ETC has long been recognised as a potential target for antimalarial chemotherapy.[113, 114] It is comprised of four enzyme complexes: NADH:ubiquinone oxidoreductase (complex I), succinate:ubiquinone oxidoreductase (complex II or SDH), ubiquinol:cytochrome c oxidoreductase (complex III or cytochrome bc_1 complex), cytochrome c oxidase (complex IV).[115] Complexes II and IV are conserved in *Plasmodia* compared to other eukaryotes, but a type II NADH dehydrogenase (*Pf*NDH2) replaces complex I. Also within the *Plasmodia* are other

oxidoreductases, such as dihydroorotate dehydrogenase (*Pf*DHODH) that is present within the mitochondria, and has an important role in *de novo* pyrimidine biosynthesis.[116] Pyrimidine biosynthesis is essential for the formation of nucleic acids, glycoproteins and phospholipids in the *Plasmodia*, as unlike many other eukaryote cells, malaria parasites obtain most of their ATP from glycolysis, rather than oxidative phosphorylation within the mitochondria.[117, 118] Glycolysis is a metabolic pathway which converts glucose into pyruvate, with the free energy released en route used to form ATP and reduced nicotinamide adenine dinucleotide (NADH). As such, the malaria parasites rely completely on pyrimidine biosynthesis for ATP generation, with the mtETC responsible for maintaining an electrochemical gradient across the mitochondrial membrane, as well as providing a constant pool of ubiquinone required for pyrimidine synthesis.[119] Thus, shutting down the mtETC completely, halts critical metabolic pathways within the microorganism, making the four enzymes within the ETC, valid and attractive targets.[120] These four drug targets have all been exploited as discovery leads for selective inhibition: *Pf*NDH2; SDH; cytochrome bc_1 complex; *Pf*DHODH.

Figure 1.13 illustrates the mtETC. *Pf*NDH2, which is the alternative complex I in *Plasmodia*, catalyses the electron transfer from NADH to ubiquinone, also known as coenzyme Q_{10} (CoQ). This maintains a constant pool of nicotinamide adenine dinucleotide (NAD^+) required for reductive metabolic pathways such as glycolysis and the tricarboxylic acid cycle,[121] also known as Krebs cycle. Krebs cycle is of central importance in all living cells, occurring in the matrix of the mitochondria and converting carbohydrates, fats and proteins into carbon dioxide and water, which are necessary to generate energy. Succinate dehydrogenase (SDH) feeds electrons along the mtETC to the cytochrome bc_1 complex, by oxidising the reduced form of flavin

adenine dinucleotide ($FADH_2$) back to flavin adenine dinucleotide (FAD), thus continuing the cycle.[122] *Pf*DHODH is the final enzyme in the biosynthesis of pyrimidine, and catalyses the oxidation of dihydroorotate to orotate at the outer side of the membrane. The pair of electrons taken from dihydroorotate in this oxidation are then transferred through the flavin mononucleotide (FMN) co-factor to ubiquinone, which has been generated at the cytochrome bc_1 complex.[123-125]

Fig. 1.13 Mitochondrial electron transport chain. (T. Rodrigues, F. Lopes and R. Moreira, *Curr. Med. Chem.*, 2010, **17**, 929-956.)

Inhibition of any of these steps interferes with the main objective of the mtETC, which is to regenerate the ubiquinone necessary for the final step of pyrimidine biosynthesis, and thus the generation of energy.[126] These enzymatic complexes are structurally different from those homologous to human enzymes, so compounds specifically designed to be potent and selective inhibitors of these *Plasmodium* enzymes promise to be exciting antimalarial agents.

*Pf*NDH2 is a relatively new drug target for which there are few known inhibitors available. What is known though it that whilst the ETC is generally well conserved across species, the first component (Complex I) differs in *P. falciparum*, and it is this difference which can be exploited for antimalarial study.[127] In eukaryotes, Complex

I is composed of NADH:quinone oxidoreductase, also known as rotenone-sensitive NADH dehydrogenase.[128] This oxidises NADH, helping to generate the electrochemical potential necessary to produce ATP. However, *P.falciparum* codes for an alternative non-proton pumping enzyme called *Pf*NDH2, also known as rotenone-insensitive NADH dehydrogenase. Inhibition of this complex stops the electron transport chain, and thus the synthesis of ATP. It is this protein which is absent in the host and specific to the parasite, making it an attractive drug target.

A known inhibitor of *Pf*NDH2 is 1-hydroxy-2-dodecyl-4(1H)-quinolone (HDQ, fig. 1.14), a highly potent compound with an IC_{50} value in the nM range. Its side effects are also limited at effective antimalarial concentrations,[129, 130] but it has been found that HDQ's specificity and effectiveness at *Pf*NDH2 are lacking, and has been found to be a more potent inhibitor of *Pf*DHODH. Thus, specific and selective inhibitors for *Pf*NDH2 have still yet to be developed. Whilst SDH is also well considered as an antimalarial drug target, many compounds which inhibit SDH, such as 5-substituted 2,3-dimethoxy-6-phytyl-1,4-benzoquinone derivatives (fig. 1.15), also have a duel inhibitory effect at *Pf*NDH2, so whilst selectivity is a concern, this is hopefully something which can be optimised in the future.[120]

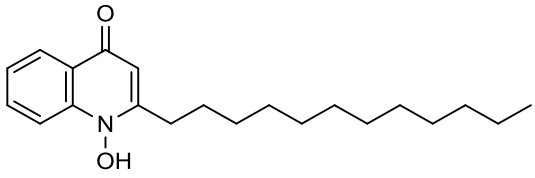


Fig. 1.14 HDQ, potent inhibitor of *Pf*NDH2.

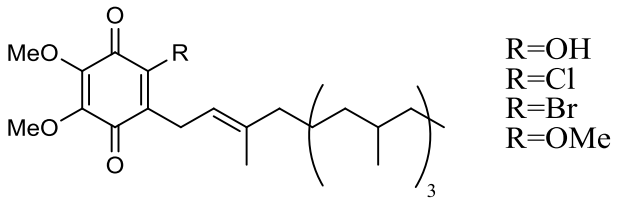


Fig. 1.15 5-substituted 2,3-dimethoxy-6-phytyl-1,4-benzoquinone derivatives with duel inhibition against SDH and *Pf*NDH2.

Whilst research is ongoing amongst the four drug targets, the most prolific and relevant to this thesis is that of complex III. This target forms the basis of much of the work described within this thesis, and will now be fully introduced.

1.3.5.1 Complex III - Cytochrome bc_1 Complex

The cytochrome bc_1 complex represents the only enzyme complex common to almost every respiratory ETC whose structures have been studied.[131, 132] In short, the cytochrome bc_1 complex is a key enzyme of the mtETC in all metazoa, as well as many fungi and protozoa. It catalyses the transfer of electrons from ubiquinol (a reduced form of CoQ) to cytochrome c, a small heme protein which transfers electrons between complex III and IV.[111, 133] This electron transfer is coupled with the translocation of protons across the inner mitochondrial membrane, with the resulting electrochemical gradient utilised for ATP generation.

The cytochrome bc_1 complex in *P. falciparum* (*Pf*bc_1) exists as a dimer of two monomeric units, each consisting of ten different polypeptides.[134] Three of these polypeptide subunits make up the catalytic core as illustrated in figure 1.16: cytochrome b; cytochrome c_1; Rieske iron-sulphur protein (ISP).[135] These three subunits participate directly in the electron transfer pathway, with evidence suggesting that the highly mobile ISP is crucial for the activity of the complex.[136-138]

The remaining subunits are likely to contribute to complex stability and the assembly process.[139]

Fig. 1.16 **(a)** Homodimeric structure of the yeast cytochrome bc_1 complex (PDB accession code 3CX5). **(b)** The structure and Q-cycle mechanism of the catalytic core of the bc_1 complex. (V. Barton, N. Fisher, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Curr. Opin. Chem. Biol.*, 2010, **14**, 440-446.)

The protomotive Q-cycle mechanism provides the most satisfactory model which accounts for the electron transfer and proton translocating activity through cytochrome bc_1, and has been extensively reviewed.[111, 133, 140-146] The Q-cycle (fig. 1.16) requires two distinct quinone binding sites; the quinol oxidation (Q_o) site and the quinone reduction (Q_i) site. These are located on the opposite sides of the membrane, linked by a transmembrane electron transfer pathway provided by the membrane spanning cytochrome b subunit. Cytochrome b provides the Q_o and Q_i sites via two B-type heme moieties bound to the subunit, termed b_L and b_H respectively. Ubiquinol that is produced by dehydrogenases, binds to the Q_o site where it is oxidised to release two protons and two electrons into the intermembrane space. In a bifurcated reaction where each electron follows down a separate path, one electron reduces the iron-sulphur cluster in the head domain of the Rieske

protein, whilst the other reduces b_L on cytochrome b. Subsequently, heme b_L is oxidised by neighbouring heme b_H, which further recycles the electron through reduction of ubiquinone to ubiquinol at the Q_i site. The reduced ISP undergoes a conformational shift in which the histidine acceptor residue at the head group rotates, allowing for close contact of ISP with the heme of cytochrome c_1, and thus the transfer of an electron. The reduced cytochrome c_1 is then oxidised by cytochrome c, which acts as an electron donor to cytochrome c oxidase (complex IV), allowing the ETC to continue.

Pfbc_1 is a major drug target in the mtETC, with several species of the complex having been cocrystallised with a number of ligands bound in the Q_o and Q_i sites, providing further insight into the complex function.[135, 147-149] *Pf*bc_1 inhibitors have a number of potential binding pockets.[120] Some compounds bind within the Q_o site, blocking the electron transfer from ubiquinol to the ISP, and electron transfer onto the b_L centre. Whilst others prevent electron transfer from ISP to cytochrome c_1, as well as electron transfer onto the b_L centre. Compounds can also inhibit the Q_i site, thereby blocking electron transfer from the b_H centre to ubiquinone.

Inhibitors specific to the Q_o or Q_i targets are well known. Myxothiazol and stigmatellin are potent Q_o inhibitors (fig. 5.2), whilst the natural antibiotic antimycin A selectively binds at Q_i (fig. 5.3). The mode of action of these compounds is well documented from a variety of crystallographic, spectroscopic and kinetic studies.[148, 150] However, the use of these compounds is limited in terms of therapeutics, as they are often highly toxic in mammals and other non-pathogenic organisms. The overall structure of the cytochrome bc_1 complex, particularly at its catalytic core, is highly conserved between species, but the Q_o site in *Pf*bc_1 does have some unusual

structural features, which may help drive efforts towards target selectivity. Selective quinol antagonists are therefore potent and attractive inhibitors.

1.3.5.2 Atovaquone

Atovaquone (ATOV, fig. 1.17) is currently the only drug in clinical use which targets *Pf*bc_1.[151-153] It inhibits the cytochrome bc_1 complex, thus collapsing the mitochondrial membrane potential leading to parasite cell death. It also displays broad antiprotozoal activity in the low nM range for several development stages of *Plasmodia*. Unfortunately, high levels of resistance to ATOV have been observed which correlate to a number of point mutations in cytochrome b. As a result of this, ATOV is now used in combination with proguanil in an attempt to improve the efficiency of the compound, and to decrease the enzymes mutation rate. In combination the drug is marketed as malarone.[154]

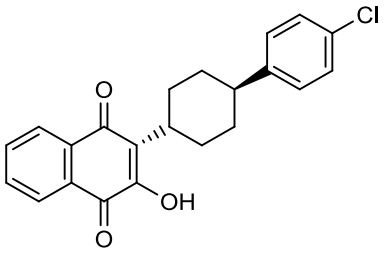


Fig. 1.17 Atovaquone, currently the only licensed drug to inhibit *Pf*bc_1.

ATOV has yet to be cocrystallised with *Pf*bc_1, so a direct study regarding its mode of action has yet to be performed. However, ATOV is a potent inhibitor of Baker's yeast (*Saccharomyces cerevisiae*), the bc_1 complex for which shows high sequence homology with that of the parasite.[155] Through electron paramagnetic resonance (EPR) spectroscopy of the Rieske ISP cluster, site-directed mutagenesis of model organism cytochrome b, and gene sequencing of ATOV resistant *Plasmodium* species, it has been shown that ATOV is a competitive inhibitor of ubiquinol at the

Q_o site.[156] This results in the collapse of the parasitic mitochondrial membrane potential, whilst having no effect on its mammalian counterpart.[119]

The Q_o site is a large domain within cytochrome b, formed from components of the C-terminal region of transmembrane helix C, the surface helix cd1 (residues 138-155), and the stretch encompassing the PEWY loop/ef helix to transmembrane helix F1 (residues 269-295).[150] ATOV binds in the Q_o site when the soluble domain of the Rieske protein is proximal to cytochrome b, and interacts directly with the ISP. This prevents mobilisation to cytochrome c_1, and consequently impairs the mitochondrial transmembrane potential.[117, 152, 157, 158].

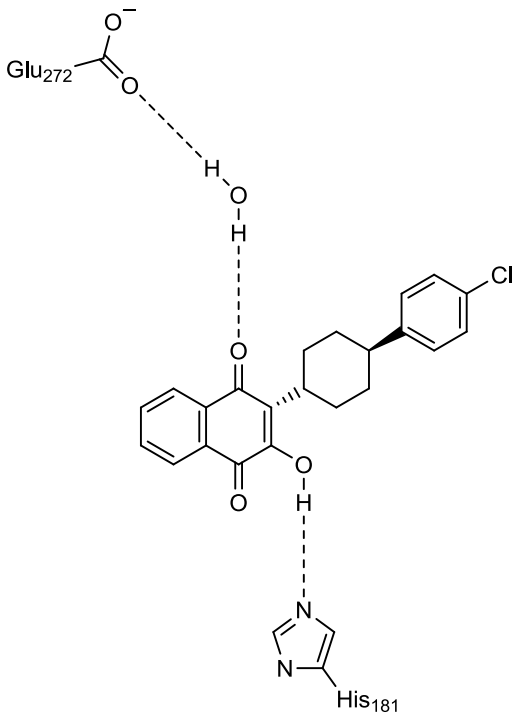


Fig. 1.18 H-bond interactions of atovaquone (X) within the Q_o site. (J. J. Kessl, B. B. Lange, T. Merbitz-Zahradnik, K. Zwicker, P. Hill, B. Meunier, H. Palsdottir, C. Hunte, S. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2003, **278**, 31312-31318.)

Figure 1.18 illustrates two hydrogen bond interactions which are observed between ATOV and amino acids in the Q_o site. The hydroxyl group of the naphthoquinone ring bonds to the imidazole nitrogen of His181 (histidine) of the Rieske protein.[159] An additional water mediated hydrogen bond interaction exists on the opposite side

of the ring system, between the carbonyl group at position 4 of the quinone ring and the carboxyl group of Glu272 (glutamic acid) of cytochrome b.[156, 157] Additional hydrophobic contacts are thought to occur between the lipophilic trans substituted *p*-chlorophenyl ring system with the side chain residues Ile147 (isoleucine) and Leu275 (leucine).[159] ATOV locks the conformation of ISP to the binding conformer, immobilising the cluster and preventing electron transfer. This collapses the membrane potential, but also impacts on the metabolic enzymes which depend on the ETC, such as *Pf*DHODH.[126]

1.3.5.2.1 Atovaquone Resistance

There has been evidence to suggest the existence of ATOV resistance in *P. falciparum* as early as the first few months of this century,[160] with clinical trials demonstrating that when used as a monotherapy, there is a high rate of recrudescence of infection.[161, 162] Resistance mutations associated with ATOV are predominantly restricted to the highly conserved PEWY region, which helps recognition and electron transfer within the Q_o site.[115] The most common point mutations observed in ATOV resistant isolates of *P. falciparum* are at codon 268 in cytochrome b.[163-166] Y268 is highly conserved across phyla and located within the ef helix of the Q_o site. The exchange of a tyrosine group in wild-type parasites to either a serine (Y268S) or asparagine (Y268N) has been found to increase the IC_{50} values of ATOV 800 to 10,000 fold.[167-169] The side chain of the residue is likely to participate in stabilising the hydrophobic interaction with bound ubiquinol,[134] with molecular modelling suggesting a similar stabilising interaction for ATOV when bound in yeast cytochrome b.[156]

However, the Y268X mutation seems not to be the only requirement for treatment failure.[167] Mutations at residue 133 from methionine to isoleucine (M133I), and at 271 from leucine to phenylalanine (L271F) have been generated *in vitro*, and have been identified in other *Plasmodium* species to bring about resistance.[166, 170, 171] Other mutations in positions 258, 267, 272 and 280 have also resulted in a 1,000 fold increase in the drugs effective IC_{50} value.[109, 172] Aside from the obvious drawbacks of emerging resistance towards ATOV, there are also other factors at play which limit its potential. Whilst the drug may have excellent antimalarial activity, it has poor pharmaceutical properties, such as low bioavailability and high plasma protein binding.[119] Despite this *Pf*bc_1 still remains a strong target for antimalarial drug development, in particular the well studied ubiquinol oxidation (Q_o) site.

Yet although there have been continued advances in the field, malaria is still one of the most prevalent and devastating diseases of our time.[173] There are relatively few effective therapies remaining, and an urgent need for novel classes of antimalarial targets and drug classes. Collaborative efforts such as the Bill & Melinda Gates foundation[174] and the Medicines for Malaria Venture (MMV)[175] now lead the way with regard to improving existing methods and developing new tools to prevent and treat malaria, with the common goal being the eradication of the disease.

1.4 Drug Design and Discovery

The drug discovery process is highly complex and requires an interdisciplinary and concerted approach to design effective and commercially viable drugs. It is a time consuming effort taking on average ten to fifteen years from preclinical discovery to regulatory approval, and whilst the total cost is subject to much contention and debate, it is now thought to cost around one billion US dollars to get a new drug to

market.[176-179] There are continued pressures to produce new and more effective therapies for a whole host of diseases, whilst simultaneously cutting down both the cost and timeframe for discovery. Experience has shown that drug discovery and development is a difficult and risky business, with only one in ten of the drugs which enter clinical development overcoming the relevant hurdles required for regulatory approval, and ultimately reaching the market.[180, 181]

The drug development process begins with the scientific study of a disease, the results of which may afford a potential target for which chemical intervention is possible. Following target identification the search can begin to find compounds which interact with this site of interest, which may be a receptor, enzyme, ion channel or even DNA or RNA.[182] Once an initial active structure has been found, a cycle of iterative steps can begin which involve optimisation and refinement of the structure, to improve its activity and pharmacological profile, ready for clinical development. If the clinical phase is successful and regulatory approval is granted, marketing can begin.

Despite a long period of success, the drug discovery pipelines have been relatively thin in the last decade, with drug launches steadily falling over recent years.[183] This may be attributed to several factors which affect the drug discovery process.[184] The more precise the medicinal target of interest is, then the less likely it is that a new drug will be developed. That is, finding a drug to treat a disease would be easier than finding a drug to target a specific part of a disease pathway. Medicinal chemists also influence the outcome of drug discovery based on their knowledge of structure optimisation, and their understanding of the disease target, as well as the facilities which are available both for synthesis and biological screening. The total cost of developing a new drug is also a hugely limiting factor during the drug discovery

process. Of about 5,000 to 10,000 compounds initially studied, only one reaches the market. Not only that but expensive synthetic routes can hamper production, as well as the high regulatory standards which must be met for new drugs.

There are a number of important features that an ideal drug should possess. It should be safe and effective at its desired target, preferably being orally absorbed and bioavailable. Metabolic stability is also important, with an attractive half-life in the body with minimal side effects and toxicity. It should also have selective distribution within the target tissues, with all of these requirements combined increasing the demands on drug discovery.

Decreasing success rates for drug discovery may be due in part to changes in conventional discovery methods. There is evidence of the use of medicines and drugs as far back as 3100 BC,[184] and for many diseases, not least that of malaria, little emphasis needs to be placed on the continued need for new drugs. However, the majority of the time, drug discovery is a process of trial and error. Commonly, drug development has revolved around the use of high throughput screening (HTS), which involves the use of robotics to quickly screen multiple compounds against a particular biochemical target. At the lead generation stage, HTS requires a library of compounds and an assay for which to measure activity of these compounds against.[185] Successful hits are generally considered to be those with IC_{50} values of about 10 micromolar (μM) or less,[186] with extensive lead optimisation typically employed to lower this value to the 1 to 10 nM range. Lead hits can also be used in the understanding of the interaction or role of a particular biochemical process. Sometimes though, HTS yields no hits,[187] and the combination of high costs and low hit rates has put the large scale approaches of the early 1990s currently out of favour.[188]

These disadvantages as well as the attraction of more deterministic approaches to combat disease led to the concept of 'rational' drug design. New understandings of the relationship between structure and biological activity have ushered in the beginning of modern drug development, with the new era cutting costs by almost a third, and development time from between 10 to 15, to 6 to 8 years.[184]

Computer aided drug design (CADD) has emerged as a powerful technique in drug discovery, and is threatening to turn more traditional approaches upside-down.[189] CADD approaches are now widely used in the pharmaceutical industry to accelerate the drug discovery process,[190, 191] and allow for more focus to be placed on the most promising compounds in a series, minimising the overall synthetic and biological testing load.

1.5 Molecular Design Loop

The molecular design loop can be considered as the 'rational' approach to most modern drug discovery endeavours, and consists of a three phase cycle of design, synthesis and testing (fig. 1.19).[192] In essence the molecular design loop involves marrying together computational efforts with chemical synthesis and biological testing in an iterative manner. Analysis of the results from one iteration provides useful information and knowledge that enables the next cycle of the loop to be initiated, with improvements made along the way. With the three specialist fields combined, chances of achieving success are greatly improved.

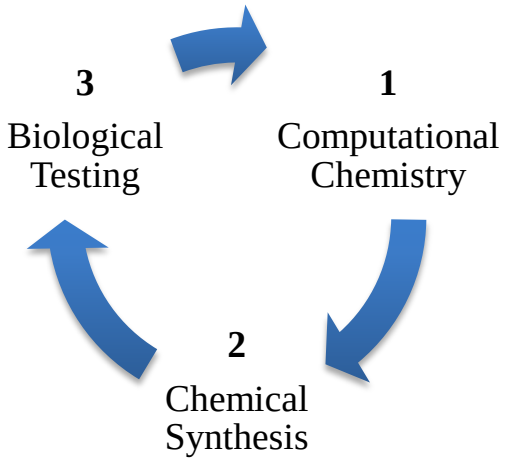


Fig. 1.19 Molecular design loop.

Whilst the focus of this thesis is very much centred on the use and application of computational chemistry techniques, it does touch on elements of chemical synthesis and biological testing. The following sections will introduce each of these topics in a context relevant to how they have been utilised.

1.6 Computational Chemistry

Computational chemistry uses the principles of computer science to assist in solving chemical problems. It is an ever expanding field which seeks to predict quantitatively, molecular structures, properties and reactivities by computational means. Modelling helps to increase predictions over conventional methods, with calculations often based on existing and readily available knowledge.[193] Its role in drug discovery is becoming more important, as it is believed to offer a means of improving efficiency to greatly reduce resource requirements, ultimately increasing the overall efficiency of drug development.[194-197] The reality is that the use of computational methods permeates all aspects of drug discovery, and those who are most proficient in its use have the advantage over their competitors.

Computational chemistry can be used in the prediction of properties related to drug-likeness, as taking onboard ADME (absorption, distribution, metabolism and excretion) consideration early in pre-clinical development, may help to avoid costly late-stage pre-clinical and clinical failures.[198] There are several protocols which already exist to try and remove compounds which appear non-drug-like before they undergo, often expensive, synthetic investigation. Possibly the most widely recognised of these is Lipinski's rule of five.[199, 200] These guidelines state that in general, orally active compounds tend to fail less than 2 of the following criteria:

- No more than five hydrogen-bond donors
- No more than ten hydrogen-bond acceptors
- Molecular weight (MW) no greater than 500 Daltons (Da)
- Partition coefficient log *P* less than five

In addition to this though, most chemical software companies now offer modules for the computation of ADME related properties,[201] the predictions of which come from relationships trained on experimental data.[202] The amount of data and reliability of the corresponding predictions does however vary, from excellent for physicochemical properties such as log *P*, to more limited for properties such as log *BB* (an index of blood-brain barrier permeability).[203] Chemical software packages such as ADMET Predictor[204] can calculate this additional information, from which informed decisions can be made as to candidates drug like potential.

1.7 Chemoinformatics

Chemoinformatics (a branch of computational chemistry) is concerned with the application of computational methods to tackle chemical problems, with particular emphasis being on the manipulation of chemical information.[192] It involves the use

of computational and informational techniques, which can be applied to a range of problems within chemistry. These *in silico* techniques are often extensively used within pharmaceutical design, with the term chemoinformatics coined in the late 1990s.[205, 206]

"Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimisation."[205, 207]

1.8 Virtual Screening

The use of chemoinformatics within the molecular design loop exist in many forms, but its widespread use to improve the efficiency of drug discovery, both in industry and academia is well documented and cannot be overstated.[186, 208-211] Most of the techniques used within chemoinformatics can be summed up under the guise of virtual screening. Virtual screening is the *in silico* analogue of biological screening, with its aim being to score, rank and/or filter a set of structures from a chemical library using one or more computational procedures, to help decide which compounds to screen, synthesise or purchase.[192] Virtual screening has been shown to be more efficient than commonly used empirical screening, with some reporting that ligand discovery hit rates (the number of compounds binding to a target, divided by the total number tested) are greater in virtual screening by two or three orders of magnitude when compared with traditional HTS.[212-215]

There are many different criteria by which structures may be scored, filtered, or otherwise assessed in a virtual screening experiment, and it has been proposed that virtual screening may be most effective when a succession of methods of increasing

complexity are used.[216] There are several main classes of virtual screening, as shown by figure 1.20. The exact method employed is dependent on the amount of structural and biological data available.[217] Methods fall within one of two categories; either they are ligand based or structure based. Structure based methods require the 3D structure of the protein drug target,[218] whilst ligand based methods can be used with only the chemical structures of known active and inactive compounds at a particular target.

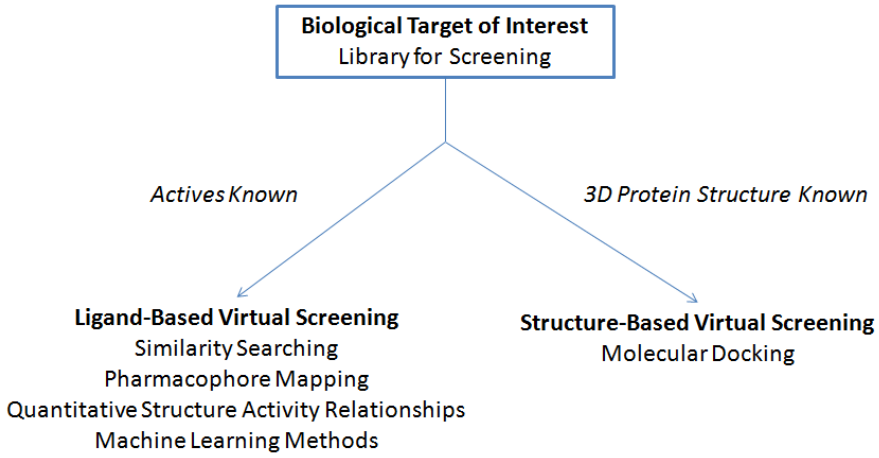


Fig. 1.20 Virtual screening techniques.

1.8.1 Ligand Based Virtual Screening

Ligand based virtual screening (LBVS) is used in the absence of a 3D protein structure, with lead identification and optimisation being dependant on the availability of pharmacologically relevant agents and their bioactivities.[219-222] Approaches include similarity searching, pharmacophore mapping, and a host of machine learning methods including quantitative structure activity relationships (QSAR).[222-224] LBVS is useful when there is little to no experimental or structural data available for a biological target of interest, with methods involving the use of biological data and the chemical structures of active/inactive compounds.[225]

1.8.1.1 Similarity Searching

Similarity searching is useful in novel hit identification when there are very few, or perhaps even only one active compound known against a particular target. If there is more than one compound available this simply acts to enrich further exploration of the chemical space which is being sampled. Similarity searching in chemical databases was first introduced in the mid 1980s,[226, 227] and offers a complementary alternative to substructure searching. Substructure searching involves the specification of a precise query structure, which is then used to search a database of compounds to identify those of interest which contain that particular substructure. This approach has its limitations though, as certainty around the chosen substructure may be ambiguous at best if only one active compound is known, and the library would simply be partitioned into those which contain the query and those which didn't. A simple query may also yield a large number of hits which are not ranked, making selection of the most promising candidates difficult.

Similarity searching on the other hand does as the name suggests. Here the query compound is used to search a database to find those compounds which are most similar to it, comparing each molecule in turn and ranking the compounds in order of decreasing similarity to the query. Similarity searching offers several advantages. For one there is no need to define a precise substructure with which to search. The user also has control over the number of compounds output, with every compound given a numerical score as a similarity measure, so a threshold can be placed only considering compounds which exceed a particular level of similarity to the query or queries.

Like many virtual screening methods, similarity searching utilises the similarity principle, which states that structurally similar compounds are more likely to exhibit similar properties.[228-231] Given a molecule of known biological activity, you would expect, according to the similarity principle, structurally similar compounds to exhibit similar activity. This characteristic has been referred to as neighbourhood behaviour.[229] Similarity searching is widely used to identify compounds for screening from a library, based on an initial molecule which is known to possess some desirable activity.

1.8.1.1.1 Tanimoto Coefficient

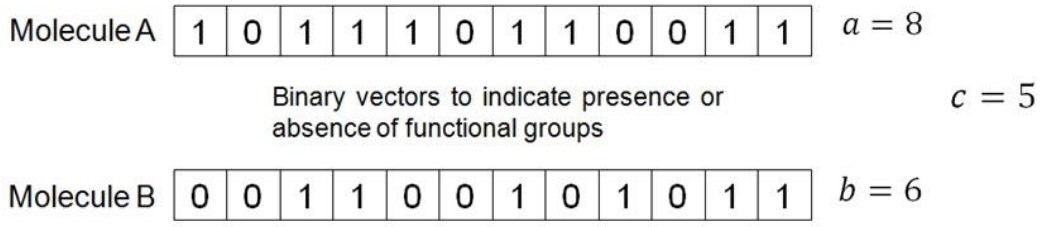
The main consideration with similarity searching is how best to assess the degree of similarity between compounds. There are many ways of quantifying similarity, but perhaps the most commonly used in virtual screening is that of the Tanimoto coefficient.[192] It is a similarity method based on two-dimensional (2D) fingerprints, which are simply binary vectors that indicate either the presence ("1") or absence ("0") of a particular substructural fragments within a molecule. It offers a method of quantifying the fragments in common between two molecules.[227] The Tanimoto coefficient between two molecules, A and B (S_{AB}), can be calculated using equation 1.1, where a represents the number of bits ("1") or fragments which are present in molecule A, b the number of bits in molecule B, and c the number of bits common to both molecules.

$$S_{AB} = \frac{c}{a + b - c}$$

Eq. 1.1 Tanimoto coefficient.

Figure 1.21 illustrates a hypothetical example of how the Tanimoto similarity coefficient is calculated. In this example, the similarity between molecules A and B

can be quantified with a Tanimoto coefficient value of 0.56. The Tanimoto coefficient can range between 0 and 1, with a value of 1 indicating that the molecules have identical fingerprint representations, and a value of 0 indicating that there is no similarity or common fragments between the two. Recent studies have suggested that 2D fingerprints generally perform better than 3D structure based methods in virtual screening.[232]



$$S_{AB} = \frac{5}{8 + 6 - 5}$$

Fig. 1.21 Calculating similarity between molecules A and B using binary vectors and the Tanimoto coefficient. (A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.)

1.8.1.2 Pharmacophore Mapping

Another LBVS approach which can be used in the absence of receptor details is pharmacophore mapping.[224] Unlike similarity searching which can be performed when only one active compound is known, pharmacophore mapping requires a number of molecules in order to build up a consensus pharmacophore model. A pharmacophore can be defined as the ensemble of steric and electronic features that are necessary to ensure the optimal supramolecular interaction with a specific biological target, to either trigger or block its response.[233, 234] Pharmacophore models are often employed and found computationally by extracting common features from the superimposed structures of known actives.[235] It does not represent a real molecule or a real association of functional groups, but instead is a purely abstract concept that accounts for the common molecular interaction capacities of a

group of compounds towards their target structure, and can be considered as the largest common denominator shared by a set of active molecules.

The pharmacophore model may represent the key positioning of pharmacophoric descriptors within a hypothetical binding site, including H-bond donors, H-bond acceptors, hydrophobic and aromatic groups, as well as positive and negative ionisable sites. These groups represent chemical features complementary to the receptor in 3D space, and will most likely be involved in binding and contribute to the activity at the target site. The use of these features is an extension of the concept of bioisosterism, which recognises that certain functional groups have similar biological, chemical and physical properties.[236, 237]

A common binding mode of all reference ligands is a prerequisite for building a meaningful model, meaning that all molecules within a dataset should be active at the same target within a biomolecular disease pathway.[225, 235, 238] There are however concerns with pharmacophore mapping, chiefly how the conformational flexibility of the molecules is taken into account, and also the potentially large number of pharmacophoric group combinations.

Pharmacophore mapping is of particular use when trying to identify structures with desirable bioactivity profiles from a large and previously unexplored area of chemical space.[239] This is due to the abstract nature of pharmacophore mapping, in that it depends on atomic properties rather than element types, and not on any specific chemical connectivity. Even a simple pharmacophore model can afford a set of considerations which can greatly reduce the search space, and provide a more focussed approach to selecting molecules from a chemical library.

1.8.1.3 Quantitative Structure Activity Relationship

Quantitative structure activity relationships (QSAR) or quantitative structure property relationships (QSPR) represents a branch of statistical machine learning used to correlate the structural or physicochemical properties of a molecule with a measured property, such as biological or chemical reactivity.[192] The idea of SAR dates back to 1868 when reports were made of the correlation between paralysing activity and the nature of quarternary groups in a collection of strychnine like compounds.[240] More recently however, studies in the 1960's by Hansch demonstrated QSARs applicability and usefulness, leading to its growing use.[241, 242] QSAR is commonly used in predictive toxicology to avoid drug attrition,[233] as toxicity has been one of the main causes of drug failure.[180] QSAR methods have therefore gone some way to answer the call for *in silico* methods for the predictive evaluation of drug toxicity, in order to minimise animal testing.

QSAR can be used to build mathematical relationships between the structural attributes and target properties of compounds in a chemical dataset.[223, 243, 244] These structural attributes or molecular descriptors as they are commonly referred to, describe the various chemical properties of the compounds, and when numerically expressed, may predict the biological activity of the compounds in the dataset. When a successful relationship has been found, it may be used to predict the biological activity of external structures not used during model generation. Such QSAR models have the general form shown in equation 1.2.

$$\Delta Activity = f(\Delta Molecular\ descriptors)$$

Eq. 1.2 General QSAR equation.

When QSAR was first derived by Hansch for a series of structurally related compounds,[241, 245, 246] the molecular descriptors which were quantified related to the electronic and hydrophobic characteristics of the compounds. Using these, a relationship was proposed between molecular structure and the property of interest. This led to equation 1.3, where C represents the concentration of compound required to produce a standard response in a given time, $logP$ the logarithm of the molecules partition coefficient between 1-octanol and water, and σ the appropriate Hammett substitution parameter. The values of k represent coefficients of the equation.

$$\log\left(\frac{1}{C}\right) = k_1 logP + k_2\sigma + k_3$$

Eq. 1.3 Hansch QSAR equation.

A set of known ligands can provide the basis for a target specific QSAR model, with the models built used to predict the potential activity of other molecules based on their own molecular structures.[247] For example, drug activity may be predicted without knowing the nature of the binding site for that compound, through a correlation of molecular properties of the ligands and their measured activities, useful when there is no 3D protein structure available.

1.8.1.3.1 Molecular Descriptors

Molecular descriptors are a set of properties/values associated with molecular structures, that encode molecular/chemical information in numerical form.[248, 249] The concept of molecular descriptors was developed with the advent of QSAR, when very early on it was discovered that the steric, electronic and hydrophobic properties of molecules were largely responsible for drug behaviour.[250] QSAR has since evolved to encompass more than 3,000 descriptors as a means of encapsulating molecular properties.[251,246]

Molecular descriptors are generated from chemical structures using programs that can perform various feature recognition calculations to produce a series of values for each compound.[252, 253] These values are compressed representations of the chemical structure, and may consist of 0, 1, 2 and 3D molecular properties. Zero-dimensional (0D) descriptors represent dimensionless properties which are independent of molecular connectivity and conformations. These may include features such as molecular weight and atom counts,[252, 254] and are essentially features which are based solely on a compounds molecular formula.[255] One-dimensional (1D) descriptors capture chemical functionality and the fragments of a molecule, and represent molecular properties such as hydrophilic factors and bond counts.[251, 254] 2D descriptors can correspond to physicochemical and topological properties,[256] encoding for charges and topological bond distances between atoms. Finally, 3D descriptors encode for structural geometry, chirality, and the respective charges and electrostatic surfaces of molecules.[257] However, assigning 3D descriptors can be problematic, as it is necessary to predict the conformation and/or alignment of the different compounds in their active conformations, which can be difficult for highly flexible molecules.[258]

It is hoped that by combining 0, 1, 2 and 3D molecular descriptors, the majority of a compounds molecular properties can be represented numerically. However, even when 3D descriptors are omitted, QSAR analysis using just 0, 1 and 2D descriptors can still provide powerful and reliable results, together with readily interpretable models.[259] Both 2D and 3D QSAR approaches have been developed successfully,[260, 261] but their use is dependent on the molecular descriptors available, and the mathematical approaches used to establish correlation between the target property and the descriptors.

1.8.1.3.2 Multiple Linear Regression

Perhaps the most popular and widely used machine learning technique for QSAR development, from which many other approaches are based upon, is linear regression. In its simplest form it has a linear equation as shown by equation 1.4,[192, 262] where y represents the dependent variable, x the independent variable, and c some sort of constant term. The dependent variable is the property one is trying to develop a QSAR model for such as the biological activity, and the independent variable would be a molecular descriptor of some sort. The aim of linear regression is to give the smallest possible sum of squared differences between the actual dependent observations, and those predicted from the regression equation.

$$y = mx + c$$

Eq. 1.4 Simple linear equation.

Multiple linear regression (MLR) is an expansion of simple linear regression, and involves more than one independent variable. It is the traditional statistical approach for deriving QSAR models for a dataset, relating the dependent variable (y), to a number of independent variables (x) using linear equations, similar to that of equation 1.5.

$$y = m_1x_1 + m_2x_2 + \cdots m_nx_n + c$$

Eq. 1.5 Multiple linear regression equation.

The gradient (m) and constant (c) can be calculated using equations 1.6 and 1.7 respectively.

$$m = \frac{\sum_{i=1}^{N}(x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}$$

Eq. 1.6 Calculating the gradient.

$$c = \langle y \rangle - m\langle x \rangle$$

Eq. 1.7 Calculating the constant.

N represents the number of data points, and $\langle x \rangle$ and $\langle y \rangle$ are the means of the independent and dependent variables represented by equations 1.8 and 1.9 respectively.

$$\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Eq. 1.8 Independent variable.

$$\langle y \rangle = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Eq. 1.9 Dependent variable.

Before developing a QSAR it is necessary to standardise the independent variables, so that each descriptor has an equal contribution to the overall analysis. The descriptors often have substantially different numerical ranges, so it is important that they be scaled appropriately, giving each an equal chance of contributing to the overall analysis. Otherwise, a descriptor which has a large range of values will overwhelm any variation seen in a descriptor which has a smaller range of values, and thus bias the results. One such method is unit variance scaling, also known as autoscaling.[192] The variables are standardised to have a mean of zero and a standard deviation (SD) of one. This is performed by dividing each descriptor value by the standard deviation for that descriptor across all observations (molecules), leaving each scaled descriptor with a variance of one.

The quality of a regression can be assessed in a number of ways, the most common of which is to calculate the squared correlation coefficient (r^2), which looks at the goodness-of-fit for the data. r^2 estimates the proportion of the variation in the

dependent variable which is explained by the regression equation.[263] It has a value between zero and one representing the extent of the relationship between the dependent and independent variables. If there is no linear relationship then $r^2 = 0$, indicating that none of the variation in the dependent variable is explained by the independent variables. Conversely, an r^2 value of 1 corresponds to a perfect fit.

The value of r^2 is found by first calculating the total sum of squares (TSS), the explained sum of squares (ESS), and the residual sum of squares (RSS), as shown by equations 1.10, 1.11 and 1.12 respectively. In these equations y_i represents the observed dependent variable, and $y_{calc,i}$ the values calculated by feeding the relevant independent variables into the regression equation.

$$\mathrm{TSS} = \sum_{i=1}^{N} (y_i - \langle y \rangle)^2$$

Eq. 1.10 Calculation of total sum of squares (TSS).

$$\mathrm{ESS} = \sum_{i=1}^{N} \left(y_{calc,i} - \langle y \rangle\right)^2$$

Eq. 1.11 Calculation of explained sum of squares (ESS).

$$\mathrm{RSS} = \sum_{i=1}^{N} \left(y_i - y_{calc,i}\right)^2$$

Eq. 1.12 Calculation of residual sum of squares (RSS).

TSS, ESS and RSS are then all used together in the manner described in equation 1.13 to calculate r^2.

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS}$$

Eq. 1.13 Calculation of the r^2 relationship.

Several recommendations have been made about the optimum value of r^2 in order for a model to be considered to have a good fit, yet without over training the data

(perfect relationship).[264] As its value is subjective it is therefore essential to use additional statistical methods in order to truly validate a model. A general rule of thumb though is that a QSAR model shows promise if it has an r^2 value greater than about 0.7.[265]

The adjusted r^2 value (r^2_{adj}) is interpreted similar to r^2, except that it takes into consideration the number of degrees of freedom.[262] It is adjusted by dividing the residual sum of squares and total sum of squares by their respective degrees of freedom, as shown by equation 1.14, where n represents the number of observations and p the number of variables. The value of r^2_{adj} decreases if an added variable does not reduce the unexplained variance in the data.

$$r^2_{adj} = 1 - (1 - r^2).\left(\frac{n-1}{n-p-1}\right)$$

Eq. 1.14 Calculation of the adjusted r^2 (r^2_{adj}).

There are many other criteria which need to be satisfied in order to be confident in a models performance. There are also many other machine learning methods which can be employed to develop QSAR models. Further discussion and review of these approaches and validation criteria can be found in Chapter VI.

1.8.1.4 Machine Learning Methods

Machine learning encompasses a whole set of techniques which can be used in LBVS, and have become increasingly popular in drug discovery because of their emphasis on obtaining accurate predictions.[266] Machine learning could fall under the heading of data mining, and can identify relationships in large, multidimensional datasets, between the molecular structures and properties of interest. In initial virtual screening it is often better to simply consider molecules as either "active" or

"inactive", or to use activity bins of moderate sizes, such as "high", "medium" or "low", rather than actual numerical values. Using these parameters, machine learning techniques can be considered as classification methods, with the aim being to derive computational models which enable the activity class of new structures to be predicted. In essence, these models can rank chemical structures according to their chances of clinical success, greatly aiding the prioritisation of compounds for synthesis or selection from a chemical library, saving time, expense and focusing biological testing.[267]

Even with only a few reference compounds, machine learning can significantly aid LBVS.[268-270] Numerous approaches have been developed for building classification models for various responses, with these models used to predict labels for a given compound based on its structure. An important method which utilises machine learning theory is support vector machines (SVM), which can also be used for the generation of QSAR models.[271, 272] SVM has become very popular as it can produce robust SAR models, by attempting to find a boundary or hyperplane that separates two classes of compounds (active and inactive).[192] The hyperplane is positioned using examples in the training set (molecules for which biological data exists) which are known as the support vectors. Molecules in the test set (those still to be classified) are mapped onto the same feature space and their activity predicted according to which side of the hyperplane they fall on. The distance of a compound from the boundary line can be used to assign a confidence level to the prediction, such that the greater the distance from the boundary line, the higher the confidence in the prediction and vice versa. One example of its use is in the prediction of active compounds against a series of G-protein coupled receptors (GPCRs).[273] Robust and extensively validated models were generated which were able to classify compounds

as active or inactive against a number of GPCR assays. SVM is explored further in Chapter VI, with alternative machine learning methods discussed in Chapter II.

1.8.1.5 Ligand Based Virtual Screening Successes

There are several examples of LBVS being used successfully within antimalarial drug design.[274-278] One involves the analysis of two diverse sets of compounds active against the D6 and NF54 parasite strains of malaria,[279] both of which are mefloquine resistant but CQ sensitive. The molecules were collected from a number of literature sources, and statistically significant QSAR models developed and used to predict the activity of compounds structurally similar to those used in the study. The ultimate validation for these models was their ability to predict activities for these new compounds, which were in good agreement with experimental data. The mechanism of action for these two series was also discussed by analyzing the physicochemical meaning behind the molecular descriptors incorporated into the QSAR equations.[279] The results provided the first step towards the prediction of novel active compounds, with modelling leading synthetic efforts.

Other work includes the analysis of inhibitory activities for a series of analogues against *P. falciparum* and the rat protein farnesyltransferase (PFT).[275] QSARs were developed for the two enzymes in order to explore the similarities and differences between the two. The results suggested that molecules with a minimum energy arrangement and a low positively/negatively charged surface area are optimum for *P. falciparum* activity, as are those which are less hydrophobic. Conversely, a more positively/negatively charged surface area seemed to be preferred for molecules active against the rat-PFT enzyme, with results from this study used to develop analogues which were selective for the malaria parasite enzyme.

Many more instances of LBVS successes exist outside of antimalarials,[280] with one example being the identification of new lead candidates which show potent dual inhibition against phosphodiesterase (PDE) -1 and -5, for development as potential cardiovascular therapeutics.[281] The virtual screening methods, which included classification and regression tree analysis using pharmacophore descriptors, demonstrated a high predictive ability for bioactivity of new chemical compounds. Of the 19 compounds which were tested in the study, 11 had greater than 50% inhibition at 10mM, with 7 of them of interest as dual PDE1 and PDE5 inhibitors. The uses of LBVS methods is continually being investigated and compared to those of SBVS.[282]

1.8.2 Structure Based Virtual Screening

Structure based virtual screening (SBVS) has played an important role in drug discovery and development.[283-286] Its use depends on the availability of a 3D protein structure, be it through x-ray crystallography, nuclear magnetic resonance (NMR), or predicted by homology modelling (the construction of a 3D model for a protein based on its amino acid sequence).[186, 218] From these structures new ligands active against a particular target may be designed.[287] Often it involves the extensive use of molecular docking to predict the binding poses and the strength of binding for potential ligands within a protein active site.[218]

Most drugs now arise through discovery programs that begin with the identification of a biomolecular target of therapeutic value.[186] To this end, the molecular docking of chemical libraries can be used to identify lead compounds which can be optimized in the molecular design loop, through the synthesis and assaying of numerous analogues around a SAR. Crystal structure determinations for complexes of some

analogues with the biomolecular target are often possible, and can greatly inform the optimisation of lead compounds by studying drug interactions within the binding site.[288]

The possible analogues of even a simple hit can define an enormous and largely inaccessible chemical space. Chemical space is a concept which represents the space spanned by all possible molecules,[289, 290] with estimates commonly citing the size of drug like chemical space (i.e. compounds with a MW < 500 Da) to contain around 10^{63} small molecules.[291, 292] A 3D structure of a compound bound to its target can serve to restrict that chemical space to only that which will be most profitable to explore. The ability to see how an active site is configured (the nature and conformations of the amino acid side chains), particularly in the presence of an inhibitor gives us a detailed insight into the requirements of a system.[288]

1.8.2.1 Molecular Docking

The most common structure based virtual screening approach is molecular docking.[285, 293] Molecular docking techniques have been developed over many years, with its initial use geared towards methods for the detailed analysis of small numbers of molecules. However, several factors have now played a part in its move towards higher throughput structure based methods. The continued development and improvement of high performance computer hardware has provided unparalleled amounts of dedicated computing power at relatively low costs to researchers. There has also been a significant effort in expanding the development of new algorithms for molecular docking, with tools for analysing the output of such calculations enabling scientists to navigate more effectively through large quantities of generated data.

The aim of molecular docking experiments is to predict the preferred orientation of one molecule to a second, when bound to form a stable complex.[294] Protein-ligand docking is of massive importance in rational drug design to predict the binding orientation of small molecules within a protein, as knowledge of the preferred orientation of a ligand may be used to predict its strength of association or binding affinity.[295] Molecular docking attempts to find the best match/fit between a receptor and a ligand, and involves the prediction of a ligand conformation/orientation within a binding site.

A large number of methods have been proposed for molecular docking,[285, 296-298] but it essentially consists of two problems. The first is the necessity of a mechanism for exploring the space of possible protein-ligand geometries, often referred to as the poses. The second is the need to be able to score or rank these poses in order to identify the most likely binding modes for each compound in the series, and to assign a priority order to the molecules.

The difficulty with molecular docking is that it involves many degrees of freedom around possible binding orientations. The translation and rotation of one molecule relative to another involves six degrees of freedom, and there are in addition the conformational degrees of freedom of both the ligand and the protein. The solvent may also play a significant role in determining the protein-ligand geometry, and the free energy of binding, even though it is often ignored. Predictions may be possible using interactive molecular graphics programs if the binding mode is well understood, or if the ligand is an analogue of a currently available x-ray structure, but generally, manual docking is difficult when dealing with large numbers of structures and novel ligands.

Docking algorithms can be classified according to the degrees of freedom that they consider.[192] Earlier algorithms only considered the translational and rotational degrees of freedom of the protein and ligand, treating each as a rigid body. The most widely used algorithms at present allow for the ligand to fully explore its conformational degrees of freedom, but some programs also allow for very limited conformational flexibility within certain protein side chains.[293, 299]

An example of a docking algorithm is DOCK.[287, 300-302] DOCK is generally considered to have been one of the major advances in molecular docking, though its earliest version only considered rigid body docking and was designed to identify molecules with a high degree of shape complementarity to the protein binding site. Overlapping spheres of varying radii, derived from the molecular surface of the protein are used to create a negative image of the active site. Ligand atoms are then matched to the sphere centres so that the distances between the atoms equal the distances between corresponding sphere centres. This enables the ligand conformation to be orientated within the active site, which once checked to ensure there are no unacceptable steric interactions, is ready for scoring.

More recent algorithms take the ligand conformational degrees of freedom into account, and can be classified according to the way they explore conformational space. The simplest way is to initially generate a range of ligand conformations using a conformational search algorithm in the absence of the receptor, and to then dock these conformations using a rigid body algorithm.[303] Conformations may also be generated on the fly in the presence of the receptor binding site.[304] Force field energy evaluation is often used to select energetically favourable conformations, as it has been found that ligands generally prefer to adopt local minimum conformation when binding to proteins.[305] This supports the use of local minimum conformations

for docking, but also suggests caution is still necessary to prevent excessive energy minimisation when generating ligand conformations for virtual screening. There are however, other methods which explore the orientational and conformational degrees of freedom at the same time. These fall under the categories of Monte Carlo algorithms, genetic algorithms and incremental construction approaches.

Monte Carlo docking algorithms are closely related to those employed for conformational search analysis.[306] At each iteration of the procedure either an internal conformation of the ligand is changed, or the entire molecule is subjected to a translation or a rotation within the binding site. If the energy of the new conformation is lower than that of its predecessor, then the new configuration is accepted. The first docking program to implement a Monte Carlo simulated annealing algorithm was that of AutoDock.[306, 307]

Genetic algorithms (GA) belong to the larger class of evolutionary algorithms (EA).[308] They are based on various computational models of Darwinian evolution, and have been widely used in computational chemistry.[309-311] GA can be used to perform molecular docking,[312-315] in which each chromosome encodes one conformation of the ligand together with its orientation within the binding site. A scoring function is then used to calculate the fitness of each member of the population, and to select individuals for further iteration. DOCK[314] and GOLD[316] are both examples of programs which have implemented GA in docking.[307] Owing to the random nature of both GA and Monte Carlo methods, it is usual to perform a number of runs to optimise the solutions, and select the structures with the highest scores.

Incremental construction approaches construct conformations of the ligand within the binding site in a series of stages.[316-318] Typically the algorithm will identify one or more base fragments which are docked into the binding site. These fragments are often large and rigid parts of the ligand such as ring systems, and the orientation of the base fragments in the active site forms the basis of a systematic conformational analysis for the remainder of the ligand, with the protein binding site providing an additional set of constraints that can be used to fine tune the search. The docking programs DOCK 4.0[319] and FlexX[316] both utilise incremental construction algorithms.[307]

1.8.2.1.1 Scoring Functions

It is important to make the distinctions between a docking study and a SBVS experiment. Docking involves the prediction of the binding mode of individual molecules, to identify the orientation that is closest in geometry to the observed x-ray structure. Studies to evaluate the performance of docking programs using datasets derived from the Protein Data Bank (PDB), showed that when the native ligand was docked backed into the active site, they were able to correctly predict the binding geometries in more than 70% of cases.[320-322] It is not however always clear which docking program will give the best result for a particular case,[323, 324] it is therefore important to carefully consider the results from individual studies.

For a SBVS experiment, once a pose has been generated in the binding site it is necessary to score or rank that ligand, using some function related to the free energy of association between the protein and ligand in forming that intermolecular complex. There are a wide range of scoring functions available,[325] and the ability to accurately predict the potency of ligand binding within a protein is of significant

value, providing useful starting points for drug discovery.[326, 327] Once the ligands are docked the resulting interactions can be scored, giving a quantitative measure of fit quality. Scoring functions are approximate mathematical methods used to predict the strength of the non-covalent interactions between two molecules after they have been docked, also referred to as binding affinity. It is common practice to use scoring functions in protein-ligand docking,[328] but they can also be used to predict the strength of intermolecular interactions between two proteins,[329] or even between a protein and DNA.[330] Scoring functions can be grouped into three categories: force field based, empirical, and knowledge based.[331]

Force field based scoring functions may make a smooth transition to empirical scoring functions, and include methods such as GOLDScore[313, 320] (see Chapter V). The scores are estimated by summing the strength of intermolecular van der Waals and electrostatic interactions between all atoms of the two molecules in the complex. Intramolecular energies of the two binding partners are also frequently included, and since binding normally takes place in the presence of water, the desolation energies of the ligand and the protein are sometimes taken into account using implicit solvation, which is a method of representing a solvent as a continuous medium, rather than as explicit solvent molecules. Force field based scoring functions are primarily derived from force fields such as AMBER,[332] which are frequently used in molecular dynamics simulations.

Empirical scoring functions include ChemScore,[333, 334] (see Chapter V) and are derived to reproduce data of experimentally determined complex structures based on physicochemical properties. They are based on counting the number of various types of interactions between the two binding partners.[335] Counting may be based on the number of ligand and receptor atoms in contact with each other, or by calculating

the change in solvent accessible surface area in the complex compared to the uncomplexed ligand and protein. The coefficients of the scoring function are usually fitted using MLR methods, and may include contributions from hydrogen bonding, ionic interactions, lipophilic interactions and the loss of internal conformational freedom of the ligand.

Knowledge based methods rely on the idea that a sufficiently large data sample can serve to derive rules and general principles inherently stored in this knowledge database.[331, 336-339] One such scoring function is DrugScore,[338] which is used to describe the binding geometry of ligands in proteins. It is based on statistical observations of intermolecular close contacts in large 3D datasets, such that the interaction potential between each ligand-protein atom pair is calculated as a potential of mean force. The method is founded on the assumption that close intermolecular interactions between certain types of atoms or functional groups that occur more frequently than one would expect by a random distribution, are likely to be energetically favourable and therefore contribute favourably to binding affinity.[340]

With the docking of a large compound library comes the generation of a vast amount of data, comprising the predicted binding pose for each compound, along with the predicted binding affinity of that ligand at the target. It is therefore conceivable that you could choose a list of compounds to be tested based upon the rank ordering of these compounds.[341] It is well known, however, that current scoring functions used in virtual screening are often inadequate at predicting the true binding affinity of a ligand for a receptor,[342] and there is currently no universally applicable scoring function.[343] One popular strategy to attempt to overcome this is the concept of consensus scoring.[344-346] In this approach, when a given docking function is used to generate the top ranked poses for the compounds in a target receptor, other scoring

functions are used to rescore the top ranked pose for each ligand. Only those top ranked compounds common to each scoring function (consensus) are then chosen for biological testing, with this approach showing improvements in the true hit outputs from virtual screening.[347]

As there are a multitude of possible parameters which govern the operation of docking programs, it is well recommended to spend time investigating the various options available for each docking run which is performed.[192] Molecular docking and particular scoring functions are discussed further in Chapter V.

1.8.2.2 Structure Based Virtual Screening Successes

The successes of SBVS are well documented,[293, 348-350] with the computational approaches used varying widely in their methodology, performance and speed. Some are capable of providing accurate binding models, whilst others are more suitable for the fast searching of large databases.[285, 351-358] SBVS has contributed significantly to the introduction of many compounds into clinical trials, as well as led to numerous drug approvals. One such drug is dorzolamide,[359] which was introduced into the market in 1995. It is a carbonic anhydrase inhibitor which acts as a topical anti-glaucoma agent, and was the first drug which resulted directly from SBVS.[360]

Another example was the discovery of compounds that inhibit DNA gyrase.[187] DNA gyrase is a well established antibacterial target, and the study initially involved the random screening of compounds using HTS. This led to no lead structures, so an alternative approach using molecular docking was considered. 350,000 compounds were docked, indicating 3,000 molecules of interest. When these 3,000 molecules were tested 150 hits were reported. 7 of these were later validated as true, novel

DNA gyrase inhibitors that bound at the ATP binding site. One compound was ten times more potent as a DNA gyrase inhibitor than novobiocin, a well known inhibitor.

A further example is work performed in developing antimicrobial agents against *Chlamydia pneumoniae*.[361] *C. pneumoniae* is an intracellular parasite that can cause pneumonia, bronchitis, sinusitis, pharyngitis, and atherosclerosis. The therapeutic target of interest was dimethyladenosine transferase, but given that no crystal structure was available for this, *Bacillus subtilis* RNA methyltransferase (PDB accession code 1QAO)[362] was used as a surrogate. A database of 300,000 compounds which had been filtered for undesirable chemical groups was docked into the protein binding site using FlexX,[316] after which the top 2,000 molecules were inspected and of these, 33 were purchased. Eight molecules demonstrated greater than 50% inhibition at 50 μM in a cell assay, demonstrating that the use of surrogate proteins is a viable option if no exact crystal structure of the target exists.[280]

SBVS can greatly reduce the drug discovery timeframe, due to an enhanced understanding of the optimal non-covalent contacts to be made.[363] However, both LBVS and SBVS techniques have the potential to enhance our knowledge and understanding as to how certain agents elicit a biological response. They can also be used to optimise existing compounds and allow for the design of novel scaffolds with greater selectivity and potency, not least in the field of antimalarial chemotherapy. Insight garnered from computational studies can be fed directly back into synthetic work, thus continuing the molecular design loop (fig. 1.19).

1.9 Chemical Synthesis

The next phase of the molecular design loop (fig 1.19), following on from computational chemistry is chemical synthesis. The importance of chemical synthesis in drug discovery needs very little emphasis, as nearly all clinical agents will have at some point undergone synthetic optimisation or design.[364]

1.9.1 Synthetic Discovery of Chloroquine

Within antimalarial chemotherapy there exists no better example of a synthetically derived drug than CQ. As previously discussed, until resistance began to emerge towards CQ it was considered the safe and affordable drug of choice in the treatment of malaria.[34] It was the serendipitous end point of efforts which began with the attempted synthesis of quinine (fig. 1.2) in 1856.[365] The total synthesis of quinine wasn't reported till much later in 1944,[366] but initial efforts lead to the synthesis of mauveine, the first synthetic chemical dye, eventually leading to the birth of the chemical industry.[367] These dyes were used to stain microorganisms to enhance their visibility under the microscope, but it was noticed that methylene blue was particularly effective in staining the malaria parasites (fig. 1.22).

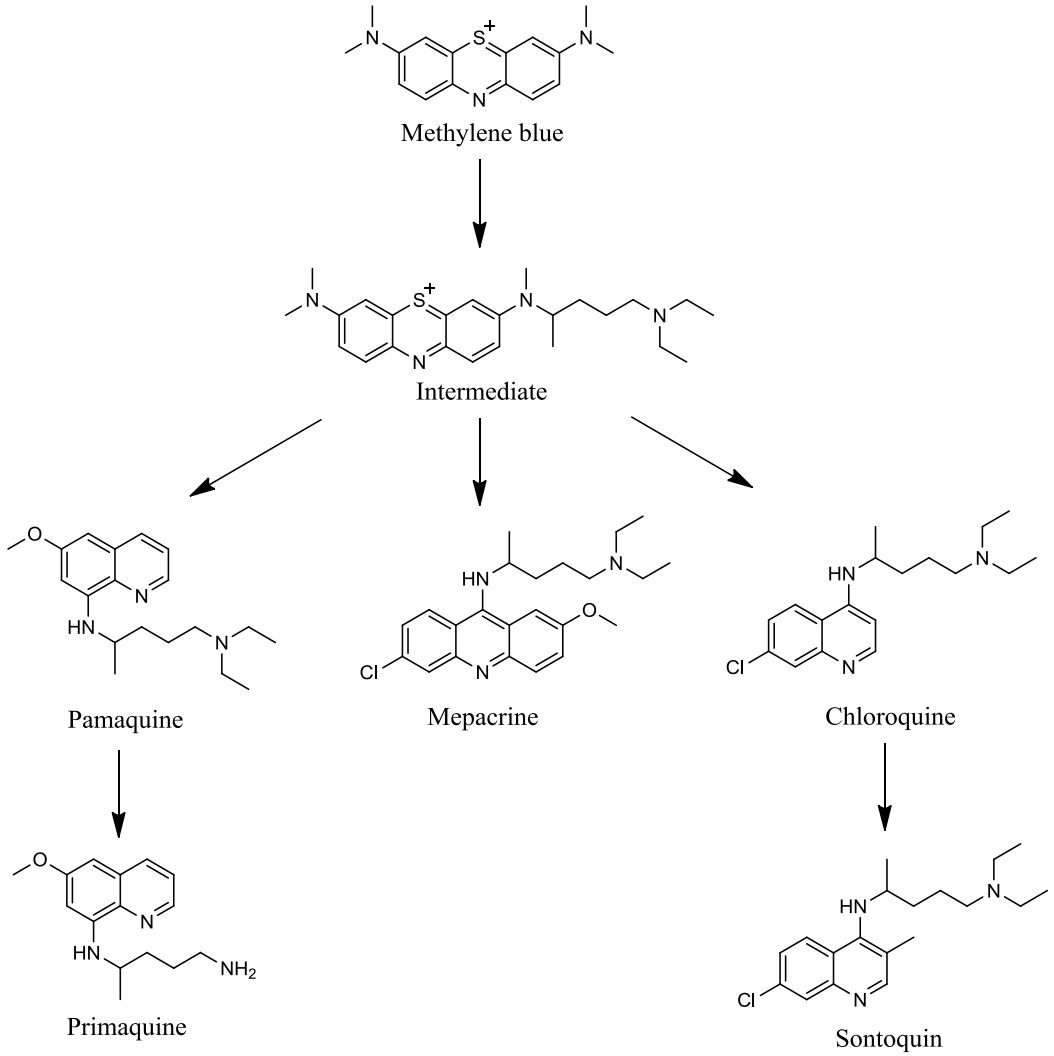


Fig. 1.22 History of synthetic efforts affording the discovery of chloroquine. (M. Schlitzer, *ChemMedChem*, 2007, **2**, 944-986.)

In 1891, two malaria patients were cured using methylene blue, and it became the first synthetic agent to be used in antimalarial therapy.[368] However, due to prominent but reversible side effects, included turning the urine green and the sclera blue, therapeutically it was not used further, but instead formed the basis of antimalarial development through chemical modifications of its structure. A key modification was the replacement of one methyl group with a dialkylaminoalkyl side chain to give an intermediate compound, the side chain of which was then connected to different heterocyclic systems such as quinoline. This gave rise to pamaquine (fig. 1.22), the first synthetic antimalarial drug, which unfortunately failed during

clinic evaluation due to multiple side effects including haemolytic anaemia.[369] The congeneric molecule primaquine was better tolerated, and became the main representative for the 8-aminoquinoline class of compounds.[370] Connection of the diethylaminoisopentylamino side chain to an acridine heterocycle yielded mepacrine, which though popular at one point when the US was cut off from its supply of quinine, had substantial side effects that included staining the eyes and skin yellow.[371] The major success in the drug design came with the introduction of the diethylaminoisopentylamino side chain into position 4 of a 7-aminoquinoline by German inventors, yielding a compound called resochin, later changed to CQ.[372] Further study also yielded sontoquin,[373] a structurally similar compound to that of CQ, but whose use was overshadowed by that of CQ, which became the foundation of antimalarial chemotherapy.[36-39]

1.9.1.1 Chloroquine Analogues

Despite its increasingly limited use, the 4-aminoquinoline chemotype of CQ is still the subject of much synthetic investigation. It is widely accepted that the 4-aminoquinoline pharmacophore plays a critical role in the complexation of CQ to FPIX to prevent the formation of haemozoin and thus parasite growth.[374] Whilst the amino groups in the side chain are considered essential for trapping high concentrations of the drug in the acidic DV of the parasite.[375]

Various studies have revealed that structural changes at the 7-position of the CQ core reduces its antimalarial activity,[375, 376] but modifications of the side chain at the 4-position pose a more promising site for optimisation. One study which detailed modifications of the side chain with *N*,*N*-diethylaminoalkyl side chains with spacers consisting of two to twelve methylene units were found to be as effective as CQ

against CQS parasite strains (fig. 1.23).[377] More importantly, the homologs with either short or long linkers between the two amino functions showed very good activity against CQR parasite strains as well.

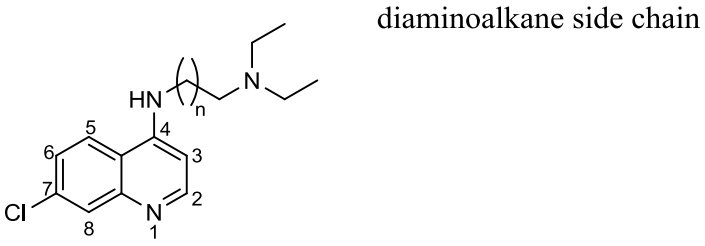


Fig. 1.23 Areas for CQ modifications.

Analogues of CQ with branched and linear side chains containing two and three methylenes between the amino groups were also found to have both *in vitro* and *in vivo* antiparasitic activity, comparable to CQ, against both CQS and CQR strains.[378, 379] More recently, work has continued to modify not only the length of the CQ side chain, but also its basicity.[380-382] In particular, success has been had with 4-amino-7-chloroquinolines which have a short linear side chain bearing two aliphatic tertiary amino functions, proving to be highly potent antimalarials of equal effectiveness against both CQS and CQR strains (fig. 1.24).[383]

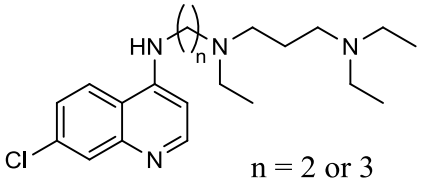


Fig. 1.24 4-amino-7-chloroquiolines with short linear side chains bearing two aliphatic tertiary amino functions for improved activity.

The modifications and studies described here clearly highlight the possibility for fresh opportunities using an old chemotype, and work will most likely continue with the aminoquinolines indefinitely. However, with many potential antimalarial targets

to hit, and an almost endless amount of chemical space to explore, it is important to consider other synthetic approaches and options available.

1.9.2 Synthesis of Novel Antimalarial Compounds

Many potential antimalarial targets have been discussed during the introduction, but perhaps the most relevant to explore with regard to the work described in this thesis (see Chapter VII), is the synthetic efforts towards *Pf*bc_1 inhibitors.

1.9.2.1 Hydroxynaphthoquinones

ATOV (fig. 1.17) is currently the only approved treatment against *Pf*bc_1, acting as a competitive inhibitor of CoQ. However, despite its excellent antimalarial activity,[119] attempts to improve its low bioavailability are ongoing, and have resulted in the design of several alternatives that substitute the 3-hydroxyl functionality for more lipophilic ester and ether groups.[384] Whilst these molecules all had potent antimalarial activity, the modifications did not improve their predicted oral bioavailabilities. In a further study, a series of ATOV alternatives were developed based on a potent hydroxynaphthoquinone inhibitor, incorporating trifluoromethyl derivatives as well as straight and branched alkyl chains onto the quinoid carbon-carbon double bond.[385] The molecule in figure 1.25 observed good activity and selectivity for the Q_o site of the parasite bc_1 complex, and these fluorinated hydroxynaphthoquinones may potentially have significant advantages over ATOV for future development.

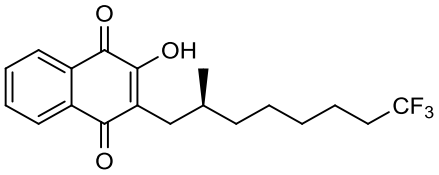
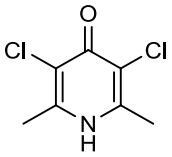


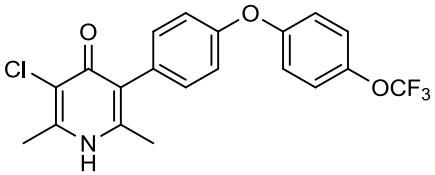
Fig. 1.25 Fluorinated hydroxynaphthoquinone derivative.

1.9.2.2 Pyridones

The antimalarial properties of pyridones such as clopidol (fig. 1.26) are well known against CQR strains of *P. falciparum*, with GlaxoSmithKline (GSK) reporting the preclinical evaluation of a new class of antimalarial 4(1H)-pyridones targeting the bc_1 complex in 2006.[386] The study found that halogenations at the C-3 position gave a ten-fold increase in activity *in vitro*, with the introduction of the ATOV *trans*-(4-chlorophenyl)cyclohexyl side chain at the C-5 position not only reducing metabolism, but also increasing *in vivo* efficacy. Substitution at C-5 with a phenoxyaryl side chain also gave increased activity,[387] with the most promising candidate being a non-chiral 4(1H)-pyridone derivative, GW844520 (fig. 1.26), which showed activity against ATOV resistant parasite strains and high selectivity for *Pf*bc_1 over mammalian bc_1.[150] GW844520 showed much promise as a drug candidate with a good half life for short term therapy, a relatively easy chemical synthetic route, and no observed cross resistance before entering preclinical development.[386] However, development of GW844520 has since been terminated owing to unexpected cardiotoxicity, which may be related to off target inhibition of human bc_1 function.[388]



Clopidol



GW844520

Fig. 1.26 Pyridone derivatives clopidol and GW844520.

Whilst GW844520 development has halted, there is still much synthetic work being performed to optimise the SAR around pyridone,[387, 389, 390] as not only do they show promising *in vivo* activity against ATOV and CQ resistant malaria strains, but they also show *in vitro* and *in vivo* activity against liver stages of the parasite, for potential use in casual antimalarial prophylaxis.

1.9.2.3 Quinolones

Quinolones are another class of antimalarial compound which act through inhibition of complex III of the ETC. They have been shown to bind at the Q_o site of the cytochrome bc_1 complex,[155] through investigation of several alkyl and alkoxy 4(1H)-quinolone derivatives of the basic core structure.[391] Simple quinolones without long side chains were found to have higher IC_{50} values than their alkyl substituted counterparts (fig. 1.27).

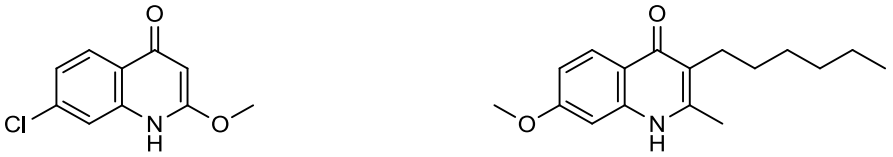


Fig. 1.27 Simple and alkyl substituted quinolones.

A trifluoromethyl head group on the terminal end of the alkyl chain resulted in a 70-fold increase in activity and was also expected to block cytochrome P_{450} mediated oxidation of the compound. However, there are concerns with these compounds

over their selectivity for *Pf*bc_1, as the flexibility of the side chains may result in off target mammalian bc_1 inhibition.

By using knowledge gained from computational study, the process of chemical synthesis may become more streamlined due to the prior selection of favourable compounds, affording a number of time and cost saving benefits. The molecular design loop (fig, 1.19) can then be completed using biological testing techniques.

1.10 Biological Testing

Testing techniques permeate all aspects of the drug discovery process,[176] and can quite often form the starting point for a drug discovery project, such as by providing the initial information required for a LBVS study.[192] Therefore, biological testing is as vital to drug discovery as any other part of the molecular design loop. Similar to chemical synthesis, the scope of biological testing greatly extends beyond that described in this thesis, and will therefore be discussed only in the context with which it has been used.

1.10.1 Bioassays

Bioassays allow for the effect of a substance against a living organism to be measured, something which is essential for the development of new drugs. They can determine either qualitatively or quantitatively, the *in vitro* effect of a substance at a particular concentration against the organism/tissue/enzyme/receptor of interest, when compared to that of a standard preparation. Bioassays have been used extensively throughout antimalarial drug design, with new assays and techniques continually being developed and optimised to yield more reliable and consistent results.[388, 392-394]

Choosing the right bioassay is crucial, as it should be quick, simple and reliable, as usually many compounds require testing.[364] During the early stages of drug discovery, *in vivo* mammalian testing is generally not possible, so testing needs to be carried out *in vitro* on isolated enzymes or membranes. Generally though, *in vitro* testing is cheaper and much easier to carry out than *in vivo* testing, with the process often being automated. Later however, if promising candidates emerge, despite the controversies surrounding human testing, *in vivo* analysis is essential to check that a drug is interacting with a specific target and having the desired pharmacological effect, as well as to monitor its pharmacokinetic properties. These properties determine the fate of a substance once it has been administered, and how it affects the body by considering its ADME properties.[21]

Target specificity and selectivity is crucial in drug discovery, as the more selective a compound is, then the less likely it is to interact with different targets and have undesirable effects. Earlier a number of biomolecular targets for malaria were discussed (i.e. *Pf*bc$_1$, *Pf*NDH2), and whilst the word antimalarial encompasses many different compounds which are active against different pathways of malaria, the problem is a lot more complex than whether they simply kill the parasite or not. An ideal target would be one which is unique to the parasite and not present in humans, to reduce the likelihood of off-target toxicity.

With increased development of resistance against many existing drug classes, antimalarial research has now moved towards more novel targets. One such example involves the study of the fourth enzyme within the pyrimidine biosynthetic pathway, *Pf*DHODH.[116, 395] HTS identified a number of chemical scaffolds which were active against malaria parasites in a whole cell growth inhibition assay, but which also observed good correlation with the *Pf*DHODH assay. That is, molecules were found

to be active against malaria, yet selective for the parasite *Pf*DHODH enzyme. This kind of approach is common within antimalarial drug design; to first identify active compounds against the malaria parasite using whole cell screening, and to then screen active hits using bioassays which determine specific sites of action.[173] Further discussion of biological testing and its various methods takes place in Chapter IV.

1.11 Aims of this Thesis

This concludes the introduction. The following chapters detail the research performed throughout this PhD. Chapter II discusses the use of LBVS methods to screen a large chemical library in order to identify novel structural chemotypes which could potentially act against malaria by inhibiting *Pf*bc_1. Chapter III outlines the complex filtering and scoring functions which were applied to enable the rational selection of the most promising candidates which resulted from LBVS, together with diversity analysis and the final selection of compounds. Chapter IV concludes the LBVS work with the testing of the selected compounds against a number of bioassays, with the resulting hits reported together with in depth interpretation of their chemical significance and possible modes of action. Chapter V details the structure based work which was performed, discussing a number of molecular docking studies which were performed to rationalise and investigate the possible mode of action of compounds which inhibit complex III of the ETC. From this work considerations were put forward with regard to optimising the activity of compounds active against *Pf*bc_1. Additionally, further support for the hits from LBVS was gathered through docking at bc_1. Chapter VI outlines a number of QSAR models which were developed for a series of 4-aminoquinoline compounds against both a

CQS and CQR strain of malaria. Additionally, a predictive model was developed to assess the drug safety of a series of thiazolide compounds active against the hepatitis C virus. Finally, Chapter VII reports a short chemical series of novel pyrroloquinolone containing compounds, which were designed, synthesised and tested for their antimalarial potential. The results from these chapters will ultimately be summarised, and the future direction of the research discussed.

1.12 References

1. WHO, *Malaria; Fact sheet Number 94*, 2009.
2. J. G. Breman, A. Egan and G. T. Keusch, *Am. J. Trop. Med. Hyg.*, 2001, **64**, IV-VII.
3. V. Patel, M. Booker, M. Kramer, L. Ross, C. A. Celatka, L. M. Kennedy, J. D. Dvorin, M. T. Duraisingh, P. Sliz, D. F. Wirth and J. Clardy, *J. Biol. Chem.*, 2008, **283**, 35078-35085.
4. S. Turschner and T. Efferth, *Mini-Rev. Med. Chem.*, 2009, **9**, 206-214.
5. R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint and S. I. Hay, *Nature*, 2005, **434**, 214-217.
6. R. M. Prothero, *Singapore Journal of Tropical Geography*, 1999, **20**, 76-85.
7. F. Castelli, S. Capone, B. Pedruzzi and A. Matteelli, *Expert Rev. Anti-Infect. Ther.*, 2007, **5**, 1031-1048.
8. M. Weill, G. Lutfalla, K. Mogensen, F. Chandre, A. Berthomieu, C. Berticat, N. Pasteur, A. Philips, P. Fort and M. Raymond, *Nature*, 2003, **423**, 136-137.
9. N. J. White, *Journal of Clinical Investigation*, 2004, **113**, 1084-1092.
10. J. Maltha and J. Jacobs, *European Journal of Pediatrics*, 2011, **170**, 821-829.
11. B. Singh, L. K. Sung, A. Matusop, A. Radhakrishnan, S. S. G. Shamsul, J. Cox-Singh, A. Thomas and D. J. Conway, *The Lancet*, 2004, **363**, 1017-1024.
12. B. Myrvang, *Tidsskrift for den Norske lageforening : tidsskrift for praktisk medicin, ny rakke*, 2010, **130**, 282-283.
13. J. Li, W. E. Collins, R. A. Wirtz, D. Rathore, A. Lal and T. F. McCutchan, *Emerg. Infect. Dis*, 2001, **7**, 35-42.
14. K. Mendis, B. J. Sina, P. Marchesini and R. Carter, *Am. J. Trop. Med. Hyg.*, 2001, **64**, 97-106.
15. F. Castelli, S. Odolini, B. Autino, E. Foca and R. Russo, *Pharmaceuticals*, 2010, **3**, 3212-3239.
16. F. Castelli, A. Matteelli, S. Caligaris, M. Gulletta, I. El-Hamad, C. Scolari, G. Chatel and G. Carosi, *Parassitologia (Rome)*, 1999, **41**, 261-265.
17. J. P. Millet, P. G. de Olalla, J. Gascon, J. G. I. Prat, B. Trevino, M. J. Pinazo, J. Cabezos, J. Munoz, F. Zarzuela and J. A. Cayla, *Malaria Journal*, 2009, **8**.
18. J. L. Gallup and J. D. Sachs, *Am. J. Trop. Med. Hyg.*, 2001, **64**, 85-96.
19. J. A. Capdevila and R. Icart, *Rev. Clin. Esp.*, 2010, **210**, 77-83.
20. D. O. Freedman, *N. Engl. J. Med.*, 2008, **359**, 603-612.
21. H. P. Rang, M. M. Dale, J. M. Ritter and P. K. Moore, *Pharmacology (Fifth Edition)*, Churchill Livingsone, 2003.
22. *Centers for Disease Control and Prevention. Available at* http://www.cdc.gov/malaria/about/biology/.
23. *Malaria Plasmodium life-cycle and natural history of malaria. Available at* http://www.malariajournal.com/sites/10007/video/plasmodium_life_cycle.html.
24. S. Pagola, P. W. Stephens, D. S. Bohle, A. D. Kosar and S. K. Madsen, *Nature*, 2000, **404**, 307-310.
25. D. J. Wyler, *Clin. Infect. Dis.*, 1993, **16**, 449-458.
26. M. Imwong, G. Snounou, S. Pukrittayakamee, N. Tanomsing, JungÂ R. Kim, A. Nandy, J. P. Guthmann, F. Nosten, J. Carlton, S. Looareesuwan, S. Nair, D. Sudimack, NicholasÂ P. Â J. Day, TimothyÂ J. Â C. Anderson and NicholasÂ J. White, *The Journal of Infectious Diseases*, 2007, **195**, 927-933.
27. F. B. Cogswell, *Clin. Microbiol. Rev.*, 1992, **5**, 26-35.
28. L. Hulden, *Malaria Journal*, 2011, **10**.
29. K. P. Krafts, E. Hempelmann and B. J. Oleksyn, *Biotech. Histochem.*, 2011, **86**, 7-35.
30. C. J. Sutherland and R. Hallett, *J. Infect. Dis.*, 2009, **199**, 1561-1563.
31. M. Amexo, R. Tolhurst, G. Barnish and I. Bates, *Lancet*, 2004, **364**, 1896-1898.
32. S. Pattanasin, S. Proux, D. Chompasuk, K. Luwiradaj, P. Jacquier, S. Looareesuwan and F. Nosten, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **97**, 672-674.
33. C. Franco-Paredes and J. I. Santos-Preciado, *Lancet Infect. Dis.*, 2006, **6**, 139-149.
34. M. Schlitzer, *ChemMedChem*, 2007, **2**, 944-986.
35. A. Kumar, S. B. Katiyar, A. Agarwal and P. M. S. Chauhan, *Curr. Med. Chem.*, 2003, **10**, 1137-1150.

36. S. R. Meshnick and M. J. Dobson, *"The History of Antimalarial Drugs" in Antimalarial Chemotherapy: Mechanisms of Action, Resistance and New Directions in Drug Discovery*, Humana, Towowa, 2001.
37. G. R. Coatney, *Am. J. Trop. Med. Hyg.*, 1963, **12**, 121-128.
38. D. Greenwood, *J. Antimicrob. Chemother.*, 1995, **36**, 857-872.
39. Schonhof.F, *Arzneimittel-Forschung*, 1965, **15**, 1256-&.
40. W. R. J. Taylor and N. J. White, *Drug Saf.*, 2004, **27**, 25-61.
41. L. Tilley, P. Loria and M. Foley, *Chloroquine and other quinoline antimalarials*, Humana Press Inc., 999 Riverview Drive, Suite 208, Totowa, NJ, 07512, USA, 2001.
42. P. M. O'Neill, S. A. Ward, N. G. Berry, J. P. Jeyadevan, G. A. Biagini, E. Asadollaly, B. K. Park and P. G. Bray, *Curr. Top. Med. Chem.*, 2006, **6**, 479-507.
43. K. Kirk and K. J. Saliba, *Drug Resistance Updates*, 2001, **4**, 335-338.
44. S. E. Francis, D. J. Sullivan and D. E. Goldberg, *Annu. Rev. Microbiol.*, 1997, **51**, 97-123.
45. P. G. Bray, S. A. Ward and P. M. O'Neill, *Curr.Top.Microbiol.Immunol.*, 2005, **295**, 3-38.
46. T. J. Egan and H. M. Marques, XXXIII International Conference on Coordination Chemistry, Florence, Italy, 1998.
47. C. D. Fitch, *Life Sci.*, 2004, **74**, 1957-1972.
48. P. G. Bray, R. E. Martin, L. Tilley, S. A. Ward, K. Kirk and D. A. Fidock, *Mol. Microbiol.*, 2005, **56**, 323-333.
49. T. N. Bennett, A. D. Kosar, L. M. B. Ursos, S. Dzekunov, A. B. S. Sidhu, D. A. Fidock and P. D. Roepe, *Mol. Biochem. Parasitol.*, 2004, **133**, 99-114.
50. K. J. Saliba, P. I. Folb and P. J. Smith, *Biochem. Pharmacol.*, 1998, **56**, 313-320.
51. T. J. Egan, *J. Inorg. Biochem.*, 2006, **100**, 916-926.
52. H. Ginsburg, S. A. Ward and P. G. Bray, *Parasitol. Today*, 1999, **15**, 357-360.
53. F. Loeb, W. M. Clark, G. R. Coatney, L. T. Coggeshall, F. R. Dieuaide, A. R. Dochez, E. G. Hakansson, E. K. Marshall, C. S. Marvel, O. R. McCoy, J. J. Sapero, W. H. Sebrell, J. A. Shannon and G. A. Carden, *Journal of the American Medical Association*, 1946, **130**, 1069-1070.
54. T. E. Wellems and C. V. Plowe, *J. Infect. Dis.*, 2001, **184**, 770-776.
55. I. M. Hastings, *Trends in Parasitology*, 2004, **20**, 512-518.
56. H. Ginsburg, *Acta Trop.*, 2005, **96**, 16-23.
57. C. P. Sanchez, J. E. McLean, W. Stein and M. Lanzer, *Biochemistry*, 2004, **43**, 16365-16373.
58. K. J. Saliba, A. M. Lehane and K. Kirk, *Mol. Microbiol.*, 2008, **70**, 775-779.
59. C. P. Sanchez, J. E. McLean, P. Rohrbach, D. A. Fidock, W. D. Stein and M. Lanzer, *Biochemistry*, 2005, **44**, 9862-9870.
60. R. E. Martin and K. Kirk, *Mol. Biol. Evol.*, 2004, **21**, 1938-1949.
61. R. A. Cooper, M. T. Ferdig, X. Z. Su, L. M. B. Ursos, J. B. Mu, T. Nomura, H. Fujioka, D. A. Fidock, P. D. Roepe and T. E. Wellems, *Mol. Pharmacol.*, 2002, **61**, 35-42.
62. D. A. Fidock, T. Nomura, A. K. Talley, R. A. Cooper, S. M. Dzekunov, M. T. Ferdig, L. M. B. Ursos, A. B. S. Sidhu, B. Naude, K. W. Deitsch, X. Z. Su, J. C. Wootton, P. D. Roepe and T. E. Wellems, *Mol. Cell.*, 2000, **6**, 861-871.
63. P. M. O'Neill, P. G. Bray, S. R. Hawley, S. A. Ward and B. K. Park, *Pharmacol. Ther.*, 1998, **77**, 29-58.
64. M. Foley and L. Tilley, *Pharmacol. Ther.*, 1998, **79**, 55-87.
65. M. D. Tingle, H. Jewell, J. L. Maggs, P. M. Oneill and B. K. Park, *Biochem. Pharmacol.*, 1995, **50**, 1113-1119.
66. D. J. Naisbitt, J. E. Ruscoe, D. Williams, P. M. Oneill, M. Pirmohamed and B. K. Park, *J. Pharmacol. Exp. Ther.*, 1997, **280**, 884-893.
67. D. J. Naisbitt, D. P. Williams, P. M. O'Neill, J. L. Maggs, D. J. Willock, M. Pirmohamed and B. K. Park, *Chem. Res. Toxicol.*, 1998, **11**, 1586-1595.
68. P. Olliaro, C. Nevill, J. LeBras, P. Ringwald, P. Mussano, P. Garner and P. Brasseur, *Lancet*, 1996, **348**, 1196-1201.
69. S. Gupta, M. M. Thapar, S. T. Mariga, W. H. Wernsdorfer and A. Bjorkman, *Exp. Parasitol.*, 2002, **100**, 28-35.
70. *http://www.who.int/malaria/publications/atoz/9789241547925/en/index.html*.
71. R. K. Haynes and S. C. Vonwiller, *Accounts of Chemical Research*, 1997, **30**, 73-79.
72. M. Ramharter, H. Noedl, K. Thimasarn, G. Wiedermann, G. Wernsdorfer and W. H. Wernsdorfer, *Am. J. Trop. Med. Hyg.*, 2002, **67**, 39-43.
73. P. Tanariya, P. Tippawangkoso, J. Karbwang, K. Na-Bangchang and W. H. Wernsdorfer, *Br. J. Clin. Pharmacol.*, 2000, **49**, 437-444.

74. C. L. Hartwig, A. S. Rosenthal, J. Dangelo, C. E. Griffin, G. H. Posner and R. A. Cooper, *Biochem. Pharmacol.*, 2009, **77**, 322-336.
75. G. H. Posner and P. M. O'Neill, *Accounts of Chemical Research*, 2004, **37**, 397-404.
76. R. K. Haynes and S. Krishna, *Microbes Infect.*, 2004, **6**, 1339-1346.
77. S. Krishna, A. C. Uhlemann and R. K. Haynes, *Drug Resistance Updates*, 2004, **7**, 233-244.
78. P. M. O'Neill and G. H. Posner, *Journal of Medicinal Chemistry*, 2004, **47**, 2945-2964.
79. K. Chotivanich, J. Sattabongkot, R. Udomsangpetch, S. Looareesuwan, N. P. J. Day, R. E. Coleman and N. J. White, *Antimicrob. Agents Chemother.*, 2006, **50**, 1927-1930.
80. N. J. White, *Antimicrob. Agents Chemother.*, 1997, **41**, 1413-1422.
81. C. J. Woodrow, R. K. Haynes and S. Krishna, *Postgrad. Med. J.*, 2005, **81**, 71-78.
82. H. Noedl, *Trends in Parasitology*, 2005, **21**, 404-405.
83. D. A. Fidock, R. T. Eastman, S. A. Ward and S. R. Meshnick, *Trends in Parasitology*, 2008, **24**, 537-544.
84. BBC, *Malaria parasites 'resist drugs'*, http://news.bbc.co.uk/1/hi/world/asia-pacific/8073118.stm.
85. H. Hugel, *Chemistry in Australia*, 2008, **75**, 7-10.
86. P. Ringwald, E. C. M. Eboumbou, J. Bickii and L. K. Basco, *Antimicrob. Agents Chemother.*, 1999, **43**, 1525-1527.
87. B. Pradines, A. Tall, T. Fusai, A. Spiegel, R. Hienne, C. Rogier, J. F. Trape, J. Le Bras and D. Parzy, *Antimicrob. Agents Chemother.*, 1999, **43**, 418-420.
88. A. Brockman, R. N. Price, M. van Vugt, D. G. Heppner, D. Walsh, P. Sookto, T. Wimonwattrawatee, S. Looareesuwan, N. J. White and F. Nosten, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 2000, **94**, 537-544.
89. E. A. Ashley and N. J. White, *Curr. Opin. Infect. Dis.*, 2005, **18**, 531-536.
90. R. K. Haynes, *Curr. Top. Med. Chem.*, 2006, **6**, 509-537.
91. M. Enserink, *SCIENCE*, 2005, **307**, 33-33.
92. P. M. O'Neill, *Expert Opinion on Investigational Drugs*, 2005, **14**, 1117-1128.
93. I. Bathurst and C. Hentschel, *Trends in Parasitology*, 2006, **22**, 301-307.
94. J. Bradbury, *Lancet Infect. Dis.*, 2004, **4**, 598-598.
95. O. Dechy-Cabaret, F. Benoit-Vical, A. Robert and B. Meunier, *Chembiochem*, 2000, **1**, 281-283.
96. L. K. Basco, O. Dechy-Cabaret, M. Ndounga, F. S. Meche, A. Robert and B. Meunier, *Antimicrob. Agents Chemother.*, 2001, **45**, 1886-1888.
97. S. A. L. Laurent, C. Loup, S. Mourgues, A. Robert and B. Meunier, *Chembiochem*, 2005, **6**, 653-658.
98. G. D. Shanks, *Milit. Med.*, 1994, **159**, 275-281.
99. A. K. Bhattacharjee and J. M. Karle, *Bioorganic & Medicinal Chemistry*, 1998, **6**, 1927-1933.
100. S. Vijaykadga, C. Rojanawatsirivej, S. Cholpol, D. Phoungmanee, A. Nakavej and C. Wongsrichanalai, *Trop. Med. Int. Health*, 2006, **11**, 211-219.
101. M. Adjuik, P. Agnamey, A. Babiker, S. Borrmann, P. Brasseur, M. Cisse, F. Cobelens, S. Diallo, J. F. Faucher, P. Garner, S. Gikunda, P. G. Kremsner, S. Krishna, B. Lell, M. Loolpapit, P. B. Matsiegui, M. A. Missinou, J. Mwanza, F. Ntoumi, P. Olliaro, P. Osimbo, P. Rezbach, E. Some and W. R. J. Taylor, *Lancet*, 2002, **359**, 1365-1372.
102. P. G. Kremsner and S. Krishna, *Lancet*, 2004, **364**, 285-294.
103. M. Adjuik, P. Agnamey, A. Babiker, J. Baptista, S. Borrmann, P. Brasseur, P. Carnevale, M. Cisse, R. Collins, U. D'Alessandro, N. Day, W. de Boom, T. Doherty, G. Dorsey, P. Garner, S. Gikunda, V. Gil, B. Greenwood, J. P. Guthmann, M. C. Henry, M. R. Kamya, P. G. Kremsner, E. Konate, S. Krishna, D. Lalloo, P. Lange, M. Loolpapit, G. Malenga, W. Marquino, K. Marsh, P. Milligan, M. Molyneux, K. Mugittu, J. Niangue, F. Nosten, F. Ntoumi, C. Obonyo, F. Ochieng, P. Olliaro, A. J. Oloo, L. Osorio, L. Pinoges, G. Priotto, P. J. Rosenthal, T. Ruebush, J. Simpson, S. Sirima, E. Some, W. Taylor, F. ter Kuile, A. Tiono, L. von Seidlein, B. Watkins, N. White and G. Int Artemisinin Study, *Lancet*, 2004, **363**, 9-17.
104. F. Grellepois, P. Grellier, D. Bonnet-Delpon and J. P. Begue, *Chembiochem*, 2005, **6**, 648-652.
105. D. C. M. Chan and A. C. Anderson, *Curr. Med. Chem.*, 2006, **13**, 377-398.
106. A. Nzila, *J. Antimicrob. Chemother.*, 2006, **57**, 1043-1054.

107. J. Yuvaniyama, P. Chitnumsub, S. Kamchonwongpaisan, J. Vanichtanankul, W. Sirawaraporn, P. Taylor, M. D. Walkinshaw and Y. Yuthavong, *Nat. Struct. Biol.*, 2003, **10**, 357-365.
108. P. K. Rathod and M. A. Phillips, *Nat. Struct. Biol.*, 2003, **10**, 316-318.
109. I. K. Srivastava, J. M. Morrisey, E. Darrouzet, F. Daldal and A. B. Vaidya, *Mol. Microbiol.*, 1999, **33**, 704-711.
110. T. Triglia, P. Wang, P. F. G. Sims, J. E. Hyde and A. F. Cowman, *Embo J.*, 1998, **17**, 3807-3815.
111. A. R. Crofts, *Annu. Rev. Physiol.*, 2004, **66**, 689-733.
112. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
113. C. C. Wang, *Journal of Medicinal Chemistry*, 1984, **27**, 1-9.
114. P. L. Olliaro and Y. Yuthavong, *Pharmacol. Ther.*, 1999, **81**, 91-110.
115. M. W. Mather, K. W. Henry and A. B. Vaidya, *Curr. Drug Targets*, 2007, **8**, 49-60.
116. M. A. Phillips and P. K. Rathod, *Infectious Disorders - Drug Targets*, 2010, **10**, 226-239.
117. M. W. Mather and A. B. Vaidya, *J. Bioenerg. Biomembr.*, 2008, **40**, 425-433.
118. M. Basselin, S. M. Hunt, H. Abdala-Valencia and E. S. Kaneshiro, *Eukaryot. Cell*, 2005, **4**, 1483-1492.
119. I. K. Srivastava, H. Rottenberg and A. B. Vaidya, *J. Biol. Chem.*, 1997, **272**, 3961-3966.
120. T. Rodrigues, F. Lopes and R. Moreira, *Curr. Med. Chem.*, 2010, **17**, 929-956.
121. A. Eschemann, A. Galkin, W. Oettmeier, U. Brandt and S. Kerscher, *J. Biol. Chem.*, 2005, **280**, 3138-3142.
122. N. Suraveratum, S. R. Krungkrai, P. Leangaramgul, P. Prapunwattana and J. Krungkrai, *Mol. Biochem. Parasitol.*, 2000, **105**, 215-222.
123. R. I. Christopherson, S. D. Lyons and P. K. Wilson, *Accounts of Chemical Research*, 2002, **35**, 961-971.
124. W. E. Gutteridge, D. Dave and W. H. G. Richards, *Biochimica Et Biophysica Acta*, 1979, **582**, 390-401.
125. J. Baldwin, A. M. Farajallah, N. A. Malmquist, P. K. Rathod and M. A. Phillips, *J. Biol. Chem.*, 2002, **277**, 41827-41834.
126. H. J. Painter, J. M. Morrisey, M. W. Mather and A. B. Vaidya, *Nature*, 2007, **446**, 88-91.
127. C. K. Dong, V. Patel, J. C. Yang, J. D. Dvorin, M. T. Duraisingh, J. Clardy and D. F. Wirth, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 972-975.
128. A. M. P. Melo, T. M. Bandeiras and M. Teixeira, *Microbiol. Mol. Biol. Rev.*, 2004, **68**, 603-+.
129. A. Saleh, J. Friesen, S. Baumeister, U. Gross and W. Bohne, *Antimicrob. Agents Chemother.*, 2007, **51**, 1217-1222.
130. S. S. Lin, S. Kerscher, A. Saleh, U. Brandt, U. Gross and W. Bohne, *Biochim. Biophys. Acta-Bioenerg.*, 2008, **1777**, 1455-1462.
131. M. Schutz, M. Brugna, E. Lebrun, F. Baymann, R. Huber, K. O. Stetter, G. Hauska, R. Toci, D. Lemesle-Meunier, P. Tron, C. Schmidt and W. Nitschke, *Journal of Molecular Biology*, 2000, **300**, 663-675.
132. D. Lemeslemeunier, P. Brivetchevillotte, J. P. Dirago, P. P. Slonimski, C. Bruel, T. Tron and N. Forget, *J. Biol. Chem.*, 1993, **268**, 15626-15632.
133. C. Hunte, H. Palsdottir and B. L. Trumpower, *FEBS Lett.*, 2003, **545**, 39-46.
134. V. Barton, N. Fisher, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Curr. Opin. Chem. Biol.*, 2010, **14**, 440-446.
135. X. G. Gao, X. L. Wen, C. A. Yu, L. Esser, S. Tsao, B. Quinn, L. Zhang, L. Yu and D. Xia, *Biochemistry*, 2002, **41**, 11692-11702.
136. E. Darrouzet, M. Valkova-Valchanova, C. C. Moser, P. L. Dutton and F. Daldal, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 4567-4572.
137. N. Fisher, I. Bourges, P. Hill, G. Brasseur and B. Meunier, *Eur. J. Biochem.*, 2004, **271**, 1292-1298.
138. U. Brandt, U. Haase, H. Schagger and G. Vonjagow, *J. Biol. Chem.*, 1991, **266**, 19958-19964.
139. V. Zara, L. Conte and B. L. Trumpower, *Febs J.*, 2009, **276**, 1900-1914.
140. J. L. Cape, M. K. Bowman and D. M. Kramer, *Trends Plant Sci.*, 2006, **11**, 46-55.
141. S. E. Chobot, H. B. Zhang, C. C. Moser and P. L. Dutton, *J. Bioenerg. Biomembr.*, 2008, **40**, 501-507.
142. P. Mitchell, *FEBS Lett.*, 1975, **59**, 137-139.

143. P. Mitchell, *Journal of Theoretical Biology*, 1976, **62**, 327-367.
144. A. R. Crofts, S. Lhee, S. B. Crofts, J. Cheng and S. Rose, *Biochim. Biophys. Acta-Bioenerg.*, 2006, **1757**, 1019-1034.
145. B. L. Trumpower, *Biochim. Biophys. Acta-Bioenerg.*, 2002, **1555**, 166-173.
146. Z. L. Zhang, L. S. Huang, V. M. Shulmeister, Y. I. Chi, K. K. Kim, L. W. Hung, A. R. Crofts, E. A. Berry and S. H. Kim, *Nature*, 1998, **392**, 677-684.
147. C. Hunte, J. Koepke, C. Lange, T. Rossmanith and H. Michel, *Struct. Fold. Des.*, 2000, **8**, 669-684.
148. L. Esser, B. Quinn, Y. F. Li, M. Q. Zhang, M. Elberry, L. Yu, C. A. Yu and D. Xia, *Journal of Molecular Biology*, 2004, **341**, 281-302.
149. S. R. N. Solmaz and C. Hunte, *J. Biol. Chem.*, 2008, **283**, 17542-17549.
150. A. R. Crofts, B. Barquera, R. B. Gennis, R. Kuras, M. Guergova-Kuras and E. A. Berry, *Biochemistry*, 1999, **38**, 15807-15826.
151. M. Fry and M. Pudney, *Biochem. Pharmacol.*, 1992, **43**, 1545-1553.
152. M. W. Mather, E. Darrouzet, M. Valkova-Valchanova, J. W. Cooley, M. T. McIntosh, F. Daldal and A. B. Vaidya, *J. Biol. Chem.*, 2005, **280**, 27458-27465.
153. J. Krungkrai, S. R. Krungkrai, N. Suraveratum and P. Prapunwattana, *Biochem. Mol. Biol. Int.*, 1997, **42**, 1007-1014.
154. A. Farnert, J. Lindberg, P. Gil, G. Swedberg, Y. Berqvist, M. M. Thapar, N. Lindegardh, S. Berezcky and A. Bjorkman, *Br. Med. J.*, 2003, **326**, 628-629.
155. R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. Oneill, *MedChemComm*, 2012, **3**.
156. J. J. Kessl, S. R. Meshnick and B. L. Trumpower, *Trends in Parasitology*, 2007, **23**, 494-501.
157. J. J. Kessl, B. B. Lange, T. Merbitz-Zahradnik, K. Zwicker, P. Hill, B. Meunier, H. Palsdottir, C. Hunte, S. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2003, **278**, 31312-31318.
158. J. J. Kessl, N. V. Moskalev, G. W. Gribble, M. Nasr, S. R. Meshnick and B. L. Trumpower, *Biochim. Biophys. Acta-Bioenerg.*, 2007, **1767**, 319-326.
159. J. J. Kessl, P. Hill, B. B. Lange, S. R. Meshnick, B. Meunier and B. L. Trumpower, *J. Biol. Chem.*, 2004, **279**, 2817-2824.
160. A. Färnert, J. Lindberg, P. Gil, G. Swedberg, Y. Berqvist, M. M. Thapar, N. Lindegårdh, S. Berezcky and A. Björkman, *BMJ*, 2003, **326**, 628-629.
161. S. Looareesuwan, C. Viravan, H. K. Webster, D. E. Kyle and C. J. Canfield, *Am. J. Trop. Med. Hyg.*, 1996, **54**, 62-66.
162. E. Suswam, D. Kyle and N. Lang-Unnasch, *Exp. Parasitol.*, 2001, **98**, 180-187.
163. M. Korsinczky, N. H. Chen, B. Kotecka, A. Saul, K. Rieckmann and Q. Cheng, *Antimicrob. Agents Chemother.*, 2000, **44**, 2100-2108.
164. L. Musset, O. Bouchaud, S. Matheron, L. Massias and J. Le Bras, *Microbes Infect.*, 2006, **8**, 2599-2604.
165. A. Berry, A. Senescau, J. Lelievre, F. Benoit-Vical, R. Fabre, B. Marchou and J. F. Magnaval, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 2006, **100**, 986-988.
166. N. Fisher and B. Meunier, *FEMS Yeast Res.*, 2008, **8**, 183-192.
167. O. Wichmann, N. Muehlberger, T. Jelinek, M. Alifrangis, G. Peyerl-Hoffmann, M. Muhlen, M. P. Grobusch, J. Gascon, A. Matteelli, H. Laferl, Z. Bisoffi, S. Ehrhardt, J. Cuadros, C. Hatz, I. Gjorup, P. McWhinney, J. Beran, S. da Cunha, M. Schulze, H. Kollaritsch, P. Kern, G. Fry, J. Richter and I. European Network Surveillance, *J. Infect. Dis.*, 2004, **190**, 1541-1546.
168. J. J. Kessl, K. H. Ha, A. K. Merritt, B. B. Lange, P. Hill, B. Meunier, S. R. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2005, **280**, 17142-17148.
169. J. E. Siregar, D. Syafruddin, H. Matsuoka, K. Kita and S. Marzuki, *Parasitol. Int.*, 2008, **57**, 229-232.
170. B. Schwobel, M. Alifrangis, A. Salanti and T. Jelinek, *Malaria Journal*, 2003, **2**.
171. D. Syafruddin, J. E. Siregar and S. Marzuki, *Mol. Biochem. Parasitol.*, 1999, **104**, 185-194.
172. J. M. Peters, N. H. Chen, M. Gatton, M. Korsinczky, E. V. Fowler, S. Manzetti, A. Saul and Q. Cheng, *Antimicrob. Agents Chemother.*, 2002, **46**, 2435-2441.
173. J. N. Burrows, K. Chibale and T. N. C. Wells, *Curr. Top. Med. Chem.*, 2011, **11**, 1226-1254.
174. *Bill & Melinda Gates Foundation* - http://www.gatesfoundation.org/topics/Pages/malaria.aspx, Accessed 16/09/2011.

175. *Medicines for Malaria Venture - http://www.mmv.org/*, Accessed 16/09/2011.
176. P. Workman, *Curr. Pharm. Design*, 2003, **9**, 891-902.
177. D. Brown and G. Superti-Furga, *Drug Discovery Today*, 2003, **8**, 1067-1077.
178. J. A. DiMasi, R. W. Hansen and H. G. Grabowski, *J. Health Econ.*, 2003, **22**, 151-185.
179. C. P. Adams and V. V. Brantner, *Health Economics*, 2010, **19**, 130-141.
180. I. Kola and J. Landis, *Nat. Rev. Drug Discov.*, 2004, **3**, 711-715.
181. R. L. Woosley and J. Cossman, *Clin. Pharmacol. Ther.*, 2007, **81**, 129-133.
182. P. Imming, C. Sinning and A. Meyer, *Nat Rev Drug Discov*, 2006, **5**, 821-834.
183. J. Drews, *Drug Discovery Today*, 2003, **8**, 411-420.
184. A. Baldi, *Systematic Reviews in Pharmacy*, 2010, **1**, 99-105.
185. R. P. Hertzberg and A. J. Pope, *Curr. Opin. Chem. Biol.*, 2000, **4**, 445-451.
186. W. L. Jorgensen, *Science*, 2004, **303**, 1813-1818.
187. H.-J. Boehm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers, N. Meunier-Keller and F. Mueller, *Journal of Medicinal Chemistry*, 2000, **43**, 2664-2674.
188. R. Lahana, *Drug Discovery Today*, 1999, **4**, 447-448.
189. K. H. Lee, *Chinese Chemical Society*, 2003, **61**, 655-670.
190. R. Gomeni, M. Bani, C. D'Angeli, M. Corsi and A. Bye, *Eur. J. Pharm. Sci.*, 2001, **13**, 261-270.
191. A. V. Veselovsky and A. S. Ivanov, *Current Drug Targets - Infectious Disorders*, 2003, **3**, 33-40.
192. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
193. N. Kumar, B. S. Hendriks, K. A. Janes, D. de Graaf and D. A. Lauffenburger, *Drug Discovery Today*, 2006, **11**, 806-811.
194. M. M. Hann and T. I. Oprea, *Curr. Opin. Chem. Biol.*, 2004, **8**, 255-263.
195. M. Stahl, W. Guba and M. Kansy, *Drug Discovery Today*, 2006, **11**, 326-333.
196. A. N. Jain, *Current Opinion in Drug Discovery & Development*, 2004, **7**, 396-403.
197. D. N. Chin, C. E. Chuaqui and J. Singh, *Mini-Rev. Med. Chem.*, 2004, **4**, 1053-1065.
198. W. J. Egan, K. M. Merz and J. J. Baldwin, *Journal of Medicinal Chemistry*, 2000, **43**, 3867-3877.
199. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 2001, **46**, 3-26.
200. C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235-249.
201. M. S. Lesney, *American Chemical Society*, 2004, **Today's Chemistry at Work**.
202. J. Taskinen and J. Yliruusi, *Adv. Drug Deliv. Rev.*, 2003, **55**, 1163-1183.
203. W. M. Pardridge, *Drug Discovery Today*, 2004, **9**, 392-393.
204. *ADMET Predictor*, http://www.simulations-plus.com/.
205. F. K. Brown, in *Annual Reports in Medicinal Chemistry*, ed. A. B. James, Academic Press, Editon edn., 1998, vol. Volume 33, pp. 375-384.
206. F. Brown, *Current Opinion in Drug Discovery & Development*, 2005, **8**, 298-302.
207. F. Brown, *Current Opinion in Drug Discovery and Development*, 2005, **8**, 298-302.
208. T. I. Oprea and H. Matter, *Curr. Opin. Chem. Biol.*, 2004, **8**, 349-358.
209. M. Congreve, C. W. Murray and T. L. Blundell, *Drug Discovery Today*, 2005, **10**, 895-907.
210. T. Hou and X. Xu, *Curr. Pharm. Design*, 2004, **10**, 1011-1033.
211. G. Keri, L. Orfi, D. Eros, B. Hegymegi-Barakonyi, C. Szantai-Kis, Z. Horvath, F. Waczek, J. Marosfalvi, I. Szabadkai, J. Pato, Z. Greff, D. Hafenbradl, H. Daub, G. Muller, B. Klebl and A. Ullrich, *Curr. Signal Transduct. Ther.*, 2006, **1**, 67-95.
212. B. K. Shoichet, *Nature*, 2004, **432**, 862-865.
213. H. J. Boehm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers and N. Meunier-Keller, *Journal of Medicinal Chemistry*, 2000, **43**, 2664-2674.
214. T. N. Doman, S. L. McGovern, B. J. Witherbee, T. P. Kasten, R. Kurumbail, W. C. Stallings, D. T. Connolly and B. K. Shoichet, *Journal of Medicinal Chemistry*, 2002, **45**, 2213-2221.
215. A. M. Paiva, D. E. Vanderwall, J. S. Blanchard, J. W. Kozarich, J. M. Williamson and T. M. Kelly, *Biochim. Biophys. Acta-Protein Struct. Molec. Enzym.*, 2001, **1545**, 67-77.
216. P. S. Charifson and W. P. Walters, *Molecular Diversity*, 2000, **5**, 185-197.
217. D. Wilton, P. Willett, K. Lawson and G. Mullier, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 469-474.
218. S. Zhang, *Methods in molecular biology (Clifton, N.J.)*, 2011, **716**, 23-38.
219. F. L. Stahura and J. Bajorath, *Comb. Chem. High Throughput Screen*, 2004, **7**, 259-269.

220. O. Guner, O. Clement and Y. Kurogi, *Curr. Med. Chem.*, 2004, **11**, 2991-3005.
221. C. Hansch, A. Leo, S. B. Mekapati and A. Kurup, *Bioorganic & Medicinal Chemistry*, 2004, **12**, 3391-3400.
222. L. Parvu, *J. Cell. Mol. Med.*, 2003, **7**, 333-335.
223. T. Langer and G. Wolber, *Pure Appl. Chem.*, 2004, **76**, 991-996.
224. O. Dror, A. Shulman-Peleg, R. Nussinov and H. J. Wolfson, *Curr. Med. Chem.*, 2004, **11**, 71-90.
225. G. Schneider and K. H. Baringhaus, *Molecular Design - Concepts and Applications*, 2008.
226. R. E. Carhart, D. H. Smith and R. Venkataraghavan, *Journal of Chemical Information and Computer Sciences*, 1985, **25**, 64-73.
227. P. Willett, V. Winterman and D. Bawden, *Journal of Chemical Information and Computer Sciences*, 1986, **26**, 36-41.
228. G. M. Maggiora and M. A. Johnson, *INTRODUCTION TO SIMILARITY IN CHEMISTRY*, John Wiley & Sons Inc, New York, 1990.
229. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, *Journal of Medicinal Chemistry*, 1996, **39**, 3049-3059.
230. Y. C. Martin, J. L. Kofron and L. M. Traphagen, *Journal of Medicinal Chemistry*, 2002, **45**, 4350-4358.
231. A. Bender and R. C. Glen, *Org. Biomol. Chem.*, 2004, **2**, 3204-3218.
232. V. Venkatraman, V. I. Perez-Nueno, L. Mavridis and D. W. Ritchie, *Journal of Chemical Information and Modeling*, 2010, **50**, 2079-2093.
233. I. M. Kapetanovic, *Chem.-Biol. Interact.*, 2008, **171**, 165-176.
234. IUPAC.
235. G. H. Grant and W. G. Richards, *Computational Chemistry*, 1998.
236. C. W. Thornber, *Chemical Society Reviews*, 1979, **8**, 563-580.
237. G. A. Patani and E. J. LaVoie, *Chemical Reviews*, 1996, **96**, 3147-3176.
238. S. Gemma, G. Campiani, S. Butini, B. P. Joshi, G. Kukreja, S. S. Coccone, M. Bernetti, M. Persico, V. Nacci, I. Fiorini, E. Novellino, D. Taramelli, N. Basilico, S. Parapini, V. Yardley, S. Croft, S. Keller-Maerki, M. Rottmann, R. Brun, M. Coletta, S. Marini, G. Guiso, S. Caccia and C. Fattorusso, *Journal of Medicinal Chemistry*, 2008, **52**, 502.
239. R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903-911.
240. A. C. Brown and T. R. Fraser, *Journal of anatomy and physiology*, 1868, **2**, 224-242.
241. C. Hansch and T. Fujita, *Journal of the American Chemical Society*, 1964, **86**, 1616-1626.
242. C. Hansch and A. R. Steward, *Journal of Medicinal Chemistry*, 1964, **7**, 691-&.
243. R. Perkins, H. Fang, W. D. Tong and W. J. Welsh, *Environ. Toxicol. Chem.*, 2003, **22**, 1666-1679.
244. A. Tropsha and W. F. Zhang, *Curr. Pharm. Design*, 2001, **7**, 599-612.
245. C. Hansch, *Accounts of Chemical Research*, 1969, **2**, 232-239.
246. R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Wiley-VCH, 2000.
247. D. Young, *Computational Chemistry - A Practical Guide for Applying Techniques to Real World Problems*, 2001.
248. C. Bologa, T. Allu, M. Olah, M. Kappler and T. Oprea, *Journal of Computer-Aided Molecular Design*, 2005, **19**, 625-635.
249. *http://www.moleculardescriptors.eu/tutorials/tutorials.htm*.
250. C. D. Selassie, *History of Quantitative Structure-Activity Relationships*, 2003.
251. R. Todeschini, V. Consonni and M. Pavan, *DRAGON 3.0*.
252. R. Todeschini, V. Consonni, A. Mauri and M. Pavan, *DRAGON Web version*.
253. Fujitsu, *ADMEWORKS ModelBuilder*, http://www.fqs.pl/life_science/admeworks_modelbuilder.
254. S. Sharma, B. K. Sharma, S. K. Sharma, P. Singh and Y. S. Prabhakar, *European Journal of Medicinal Chemistry*, 2009, **44**, 1377-1382.
255. H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins and W. Tong, *Journal of Chemical Information and Modeling*, 2008, **48**, 1337-1344.
256. A. Givehchi, A. Bender and R. C. Glen, *Journal of Chemical Information and Modeling*, 2006, **46**, 1078-1083.
257. I. V. Svitanko, D. A. Devetyarov, D. E. Tcheboukov, M. S. Dolmat, A. M. Zakharov, S. S. Grigor'eva, V. T. Chichua, L. A. Ponomareva and M. I. Kumskov, *Mendeleev Commun.*, 2007, **17**, 90-91.
258. A. Berglund, M. C. D. Rosa and S. Wold, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 601-612.

259. M. Shen, A. LeTiran, Y. D. Xiao, A. Golbraikh, H. Kohn and A. Tropsha, *Journal of Medicinal Chemistry*, 2002, **45**, 2811-2823.
260. A. J. Leo and C. Hansch, *Perspect. Drug Discov. Design*, 1999, **17**, 1-25.
261. R. Garg, A. Kurup, S. B. Mekapati and C. Hansch, *Bioorganic & Medicinal Chemistry*, 2003, **11**, 621-628.
262. OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, Paris, 2007.
263. D. L. Massart, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier Science, 1997.
264. H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry*, VCH, Weinheim, 1993.
265. Y. Li, J. Liu, D. Pan and A. J. Hopfinger, *Toxicological Sciences*, 2005, **88**, 434-446.
266. N. Wale, *Drug Development Research*, 2011, **72**, 112-119.
267. S. Agarwal, D. Dugar and S. Sengupta, *Journal of Chemical Information and Modeling*, 2010, **50**, 716-731.
268. B. Chen, R. F. Harrison, G. Papadatos, P. Willett, D. J. Wood, X. Q. Lewell, P. Greenidge and N. Stiefl, *Journal of Computer-Aided Molecular Design*, 2007, **21**, 53-62.
269. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *Journal of Chemical Information and Modeling*, 2006, **46**, 462-470.
270. I. Muegge and S. Oloff, *Drug Discovery Today: Technologies*, 2006, **3**, 405-411.
271. V. Vapnik, *The Support Vector method of function estimation*, Kluwer Academic Publishers, Norwell, 1998.
272. N. Wale, I. A. Watson and G. Karypis, *Knowl. Inf. Syst.*, 2008, **14**, 347-375.
273. J. C. Saeh, P. D. Lyne, B. K. Takasaki and D. A. Cosgrove, *Journal of Chemical Information and Modeling*, 2005, **45**, 1122-1133.
274. M. Grigorov, J. Weber, J. M. J. Tronchet, C. W. Jefford, W. K. Milhous and D. Maric, *Journal of Chemical Information and Computer Sciences*, 1997, **37**, 124.
275. M. K. Gupta and Y. S. Prabhakar, *European Journal of Medicinal Chemistry*, 2008, **43**, 2751-2767.
276. F. J. B. Cardoso, A. F. de Figueiredo, M. D. S. Lobato, R. M. de Miranda, R. C. O. de Almeida and J. C. Pinheiro, *J. Mol. Model.*, 2008, **14**, 39-48.
277. X. Gironés, A. Gallegos and R. Carbó-Dorca, *Journal of Computer-Aided Molecular Design*, 2001, **15**, 1053-1063.
278. N. Mahmoudi, J. V. de Julian-Ortiz, L. Ciceron, J. Galvez, D. Mazier, M. Danis, F. Derouin and R. Garcia-Domenech, *J. Antimicrob. Chemother.*, 2006, **57**, 489-497.
279. A. R. Katritzky, O. V. Kulshyn, I. Stoyanova-Slavova, D. A. Dobehev, M. Kuanar, D. C. Fara and M. Karelson, *Bioorganic & Medicinal Chemistry*, 2006, **14**, 2333-2357.
280. J. A. Bikker and L. S. Narasimhan, Editon edn., 2010, vol. 5, pp. 85-124.
281. K. Yamazaki, N. Kusunose, K. Fujita, H. Sato, S. Asano, A. Dan and M. Kanaoka, *Bioorg. Med. Chem. Lett.*, 2006, **16**, 1371-1379.
282. P. Tiikkainen, P. Markt, G. Wolber, J. Kirchmair, S. Distinto, A. Poso and O. Kallioniemi, *Journal of Chemical Information and Modeling*, 2009, **49**, 2168-2178.
283. H. Kubinyi, *Nat. Rev. Drug Discov.*, 2003, **2**, 665-668.
284. M. R. Reddy and M. D. Erion, *J. Enzym. Inhib.*, 1998, **14**, 1-14.
285. R. D. Taylor, P. J. Jewsbury and J. W. Essex, *Journal of Computer-Aided Molecular Design*, 2002, **16**, 151-166.
286. S. Kalyaanamoorthy and Y.-P. P. Chen, *Drug Discovery Today*, 2011, **16**, 831-839.
287. I. D. Kuntz, E. C. Meng and B. K. Shoichet, *Accounts of Chemical Research*, 1994, **27**, 117-123.
288. D. M. F. van Aalten, K. G. Milne, J. Y. Zou, G. J. Kleywegt, T. Bergfors, M. A. J. Ferguson, J. Knudsen and T. A. Jones, *Journal of Molecular Biology*, 2001, **309**, 181-192.
289. C. Mehlin, *Comb. Chem. High Throughput Screen*, 2005, **8**, 5-14.
290. C. M. Dobson, *Nature*, 2004, **432**, 824-828.
291. J. E. Peironcely, T. Reijmers, L. Coulier, A. Bender and T. Hankemeier, *PLoS ONE*, 2011, **6**.
292. R. S. Bohacek, C. McMartin and W. C. Guida, *Med. Res. Rev.*, 1996, **16**, 3-50.
293. B. K. Shoichet, S. L. McGovern, B. Q. Wei and J. J. Irwin, *Curr. Opin. Chem. Biol.*, 2002, **6**, 439-446.
294. T. Lengauer and M. Rarey, *Curr. Opin. Struct. Biol.*, 1996, **6**, 402-406.
295. D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat Rev Drug Discov*, 2004, **3**, 935-949.

296. J. M. Blaney and J. S. Dixon, *Perspect. Drug Discov. Design*, 1993, **1**, 301-319.
297. R. Abagyan and M. Totrov, *Curr. Opin. Chem. Biol.*, 2001, **5**, 375-382.
298. I. Halperin, B. Y. Ma, H. Wolfson and R. Nussinov, *Proteins*, 2002, **47**, 409-443.
299. H. A. Carlson, *Curr. Opin. Chem. Biol.*, 2002, **6**, 447-452.
300. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, *Journal of Molecular Biology*, 1982, **161**, 269-288.
301. I. D. Kuntz, *SCIENCE*, 1992, **257**, 1078-1082.
302. R. L. Desjarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz and R. Venkataraghavan, *Journal of Medicinal Chemistry*, 1988, **31**, 722-729.
303. S. K. Kearsley, D. J. Underwood, R. P. Sheridan and M. D. Miller, *Journal of Computer-Aided Molecular Design*, 1994, **8**, 565-582.
304. R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *Journal of Medicinal Chemistry*, 2004, **47**, 1739-1749.
305. Q. Wang and Y.-P. Pang, *PLoS ONE*, 2007, **2**, e820.
306. D. S. Goodsell and A. J. Olson, *Proteins*, 1990, **8**, 195-202.
307. Z. Zsoldos, D. Reid, A. Simon, B. S. Sadjad and A. P. Johnson, *Current Protein and Peptide Science*, 2006, **7**, 421-435.
308. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1989.
309. D. E. Clark and D. R. Westhead, *Journal of Computer-Aided Molecular Design*, 1996, **10**, 337-358.
310. G. Jones, *Genetic and Evolutionary Algorithms*, Encyclopedia of Computational Chemistry. Wiley, Chichester, 1998.
311. D. E. Clark, *Evolutionary Algorithms in Molecular Design*, Weinheim, Wiley-VCH, 2000.
312. R. S. Judson, E. P. Jaeger and A. M. Treasurywala, *Theochem-J. Mol. Struct.*, 1994, **114**, 191-206.
313. G. Jones, P. Willett and R. C. Glen, *Journal of Molecular Biology*, 1995, **245**, 43-53.
314. C. M. Oshiro, I. D. Kuntz and J. S. Dixon, *Journal of Computer-Aided Molecular Design*, 1995, **9**, 113-130.
315. D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel and S. T. Freer, *Chem. Biol.*, 1995, **2**, 317-324.
316. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *Journal of Molecular Biology*, 1996, **261**, 470-489.
317. A. R. Leach and I. D. Kuntz, *Journal of Computational Chemistry*, 1992, **13**, 730-748.
318. W. Welch, J. Ruppert and A. N. Jain, *Chem. Biol.*, 1996, **3**, 449-462.
319. T. J. A. Ewing, S. Makino, A. G. Skillman and I. D. Kuntz, *Journal of Computer-Aided Molecular Design*, 2001, **15**, 411-428.
320. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *Journal of Molecular Biology*, 1997, **267**, 727-748.
321. B. Kramer, M. Rarey and T. Lengauer, *Proteins*, 1999, **37**, 228-241.
322. J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins*, 2002, **49**, 457-471.
323. M. Kontoyianni, L. M. McClellan and G. S. Sokol, *Journal of Medicinal Chemistry*, 2004, **47**, 558-565.
324. G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, *Journal of Medicinal Chemistry*, 2006, **49**, 5912-5931.
325. Ajay and M. A. Murcko, *Journal of Medicinal Chemistry*, 1995, **38**, 4953-4967.
326. R. Rajamani and A. C. Good, *Current Opinion in Drug Discovery and Development*, 2007, **10**, 308-315.
327. M. H. J. Seifert, J. Kraus and B. Kramer, *Current Opinion in Drug Discovery and Development*, 2007, **10**, 298-307.
328. A. N. Jain, *Current Protein and Peptide Science*, 2006, **7**, 407-420.
329. M. F. Lensink, R. Méndez and S. J. Wodak, *Proteins: Structure, Function, and Bioinformatics*, 2007, **69**, 704-718.
330. T. A. Robertson and G. Varani, *Proteins: Structure, Function, and Bioinformatics*, 2007, **66**, 359-374.
331. O. Korb, T. Stützle and T. E. Exner, *Journal of Chemical Information and Modeling*, 2009, **49**, 84-96.
332. P. K. Weiner and P. A. Kollman, *Journal of Computational Chemistry*, 1981, **2**, 287-303.

333. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 425-445.
334. C. W. Murray, T. R. Auton and M. D. Eldridge, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 503-519.
335. H. J. Böhm, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 309-323.
336. I. Muegge and Y. C. Martin, *Journal of Medicinal Chemistry*, 1999, **42**, 791-804.
337. J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster and J. M. Thornton, *Journal of Computational Chemistry*, 1999, **20**, 1177-1185.
338. H. Gohlke, M. Hendlich and G. Klebe, *Journal of Molecular Biology*, 2000, **295**, 337-356.
339. C. A. Sotriffer, H. Gohlke and G. Klebe, *Journal of Medicinal Chemistry*, 2002, **45**, 1967-1970.
340. I. Muegge, *Journal of Medicinal Chemistry*, 2005, **49**, 5895-5902.
341. P. D. Lyne, *Drug Discovery Today*, 2002, **7**, 1047-1055.
342. D. A. Pearlman and P. S. Charifson, *Journal of Medicinal Chemistry*, 2001, **44**, 3417-3423.
343. C. Bissantz, G. Folkers and D. Rognan, *Journal of Medicinal Chemistry*, 2000, **43**, 4759-4767.
344. P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *Journal of Medicinal Chemistry*, 1999, **42**, 5100-5109.
345. C. Perez and A. R. Ortiz, *Journal of Medicinal Chemistry*, 2001, **44**, 3768-3785.
346. R. Wang, Y. Lu and S. Wang, *Journal of Medicinal Chemistry*, 2003, **46**, 2287-2303.
347. M. Stahl and M. Rarey, *Journal of Medicinal Chemistry*, 2001, **44**, 1035-1042.
348. B. E. Maryanoff, *J. Med. Chem*, 2003, **47**, 769.
349. L. W. Hardy and A. Malikayil, *Current Drug Discovery*, 2003, 15-19.
350. J. F. Blake and E. R. Laird, *Annual Reports in Medicinal Chemistry, Vol 38*, 2003, **38**, 305-314.
351. L. S. Chen, B. J. Nowak, M. L. Ayres, N. L. Krett, S. T. Rosen, S. X. Zhang and V. Gandhi, *Biochem. Pharmacol.*, 2009, **78**, 583-591.
352. L. Du-Cuny, Z. H. Song, S. Moses, G. Powis, E. A. Mash, E. J. Meuillet and S. X. Zhang, *Bioorganic & Medicinal Chemistry*, 2009, **17**, 6983-6992.
353. D. Mahadevan, G. Powis, E. A. Mash, B. George, V. M. Gokhale, S. X. Zhang, K. Shakalya, L. Du-Cuny, M. Berggren, M. A. Ali, U. Jana, N. Ihle, S. Moses, C. Franklin, S. Narayan, N. Shirahatti and E. J. Meuillet, *Mol. Cancer Ther.*, 2008, **7**, 2621-2632.
354. S. A. Moses, M. A. Ali, Z. H. Song, D. C. Lei, L. L. Zhou, R. Lemos, N. Ihle, A. G. Skillman, S. X. Zhang, E. A. Mash, G. Powis and E. J. Meuillet, *Cancer Res.*, 2009, **69**, 5073-5081.
355. S. X. Zhang, W. S. Ying, T. J. Siahaan and S. D. S. Jois, *Peptides*, 2003, **24**, 827-835.
356. S. Zhang, K. Kumar, X. Jiang, A. Wallqvist and J. Reifman, *BMC Bioinformatics*, 2008, **9**.
357. S. X. Zhang, A. H. Kaplan and A. Tropsha, *Proteins-Structure Function and Bioinformatics*, 2008, **73**, 742-753.
358. S. Zhang and L. Du-Cuny, *International journal of bioinformatics research and applications*, 2009, **5**, 269-279.
359. H. Kubinyi, *Journal of Receptors and Signal Transduction*, 1999, **19**, 15-39.
360. J. Greer, J. W. Erickson, J. J. Baldwin and M. D. Varneyl, *J. Med. Chem*, 1994, **37**.
361. J. K. O. Alvesalo, A. Siiskonen, M. J. Vainio, P. S. M. Tammela and P. M. Vuorela, *Journal of Medicinal Chemistry*, 2006, **49**, 2353-2356.
362. G. Schluckebier, P. Zhong, K. D. Stewart, T. J. Kavanaugh and C. Abad-Zapatero, *Journal of Molecular Biology*, 1999, **289**, 277-291.
363. M. MacCoss and T. A. Baillie, *SCIENCE*, 2004, **303**, 1810.
364. G. L. Patrick, *An Introduction to Medicinal Chemistry*, Oxford University Press, 2005.
365. W. H. Perkin, *Journal of the Chemical Society, Transactions*, 1896, **69**, 596-637.
366. R. B. Woodward and W. E. Doering, *Journal of the American Chemical Society*, 1944, **66**, 849-849.
367. A. S. Travis, *Technology and Culture*, 1990, **31**, 51-82.
368. R. H. Schirmer, B. Coulibaly, A. Stich, M. Scheiwein, H. Merkle, J. Eubel, K. Becker, H. Becher, O. Müller, T. Zich, W. Schiek and B. Kouyaté, *Redox Report*, 2003, **8**, 272-275.
369. A. W. Sweeney, C. R. B. Blackburn and K. H. Rieckmann, *The American Journal of Tropical Medicine and Hygiene*, 2004, **71**, 187-189.
370. J. K. Baird and K. H. Rieckmann, *Trends in Parasitology*, 2003, **19**, 115-120.
371. http://www.drugs.com/mmx/quinacrine-hydrochloride.html#, *Drugs.com - Quinacrine*, Accessed 12/09/2011.

372. http://www.cdc.gov/malaria/about/history/#chloroquine, *CDC - Chloroquine*, Accessed 12/09/2011.
373. *Br. Med. J.*, 1946, **2**, 267-268.
374. S. R. Cheruku, S. Maiti, A. Dorn, B. Scorneaux, A. K. Bhattacharjee, W. Y. Ellis and J. L. Vennerstrom, *Journal of Medicinal Chemistry*, 2003, **46**, 3166-3169.
375. T. J. Egan, R. Hunter, C. H. Kaschula, H. M. Marques, A. Misplon and J. Walden, *Journal of Medicinal Chemistry*, 2000, **43**, 283-291.
376. C. H. Kaschula, T. J. Egan, R. Hunter, N. Basilico, S. Parapini, D. Taramelli, E. Pasini and D. Monti, *Journal of Medicinal Chemistry*, 2002, **45**, 3531-3539.
377. D. Y. De, F. M. Krogstad, F. B. Cogswell and D. J. Krogstad, *Am. J. Trop. Med. Hyg.*, 1996, **55**, 579-583.
378. R. G. Ridley, W. Hofheinz, H. Matile, C. Jaquet, A. Dorn, R. Masciadri, S. Jolidon, W. F. Richter, A. Guenzi, M. A. Girometta, H. Urwyler, W. Huber, S. Thaithong and W. Peters, *Antimicrob. Agents Chemother.*, 1996, **40**, 1846-1854.
379. P. A. Stocks, K. J. Raynes, P. G. Bray, B. K. Park, P. M. O'Neill and S. A. Ward, *Journal of Medicinal Chemistry*, 2002, **45**, 4975-4983.
380. D. P. Iwaniuk, E. D. Whetmore, N. Rosa, K. Ekoue-Kovi, J. Alumasa, A. C. de Dios, P. D. Roepe and C. Wolf, *Bioorganic and Medicinal Chemistry*, 2009, **17**, 6560-6566.
381. K. Ekoue-Kovi, K. Yearick, D. P. Iwaniuk, J. K. Natarajan, J. Alumasa, A. C. de Dios, P. D. Roepe and C. Wolf, *Bioorganic & Medicinal Chemistry*, 2009, **17**, 270-283.
382. J. K. Natarajan, J. N. Alumasa, K. Yearick, K. A. Ekoue-Kovi, L. B. Casabianca, A. C. de Dios, C. Wolf and P. D. Roepe, *Journal of Medicinal Chemistry*, 2008, **51**, 3466-3479.
383. K. Yearick, K. Ekoue-Kovi, D. P. Iwaniuk, J. K. Natarajan, J. Alumasa, A. C. de Dios, P. D. Roepe and C. Wolf, *Journal of Medicinal Chemistry*, 2008, **51**, 1995-1998.
384. S. El Hage, M. Ane, J. L. Stigliani, M. Marjorie, H. Vial, G. Baziard-Mouysset and M. Payard, *European Journal of Medicinal Chemistry*, 2009, **44**, 4778-4782.
385. L. M. Hughes, R. Covian, G. W. Gribble and B. L. Trumpower, *Biochim. Biophys. Acta-Bioenerg.*, 2010, **1797**, 38-43.
386. H. Xiang, J. McSurdy-Freed, G. S. Moorthy, E. Hugger, R. Bambal, C. Han, S. Ferrer, D. Gargallo and C. B. Davis, *J. Pharm. Sci.*, 2006, **95**, 2657-2672.
387. C. L. Yeates, J. F. Batchelor, E. C. Capon, N. J. Cheesman, M. Fry, A. T. Hudson, M. Pudney, H. Trimming, J. Woolven, J. M. Bueno, J. Chicharro, E. Fernandez, J. M. Fiandor, D. Gargallo-Viola, F. G. de las Heras, E. Herreros and M. L. Leon, *Journal of Medicinal Chemistry*, 2008, **51**, 2845-2852.
388. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
389. M. B. Jimenez-Diaz, T. Mulet, S. Viera, V. Gomez, H. Garuti, J. Ibanez, A. Alvarez-Doval, L. D. Shultz, A. Martinez, D. Gargallo-Viola and I. Angulo-Barturen, *Antimicrob. Agents Chemother.*, 2009, **53**, 4533-4536.
390. T. Rodrigues, R. C. Guedes, D. dos Santos, M. Carrasco, J. Gut, P. J. Rosenthal, R. Moreira and F. Lopes, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 3476-3480.
391. R. W. Winter, J. X. Kelly, M. J. Smilkstein, R. Dodean, D. Hinrichs and M. K. Riscoe, *Exp. Parasitol.*, 2008, **118**, 487-497.
392. N. Fisher, C. K. Castleden, I. Bourges, G. Brasseur, G. Dujardin and B. Meunier, *J. Biol. Chem.*, 2004, **279**, 12951-12958.
393. D. W. Wilson, B. S. Crabb and J. G. Beeson, *Malaria Journal*, 2010, **9**.
394. N. Fisher, A. J. Warman, S. A. Ward and G. A. Biagini, in *Methods in Enzymology, Vol 456*, ed. W. S. Allison, Elsevier Academic Press Inc, San Diego, Editon edn., 2009, vol. 456, pp. 303-320.
395. M. A. Phillips, R. Gujjar, N. A. Malmquist, J. White, F. El Mazouni, J. Baldwin and P. K. Rathod, *Journal of Medicinal Chemistry*, 2008, **51**, 3649-3653.

Chapter II

Ligand Based Virtual Screening Methods

2. Ligand Based Virtual Screening Methods 86

2.1 Identification of Initial Data 86

2.2 Chemical Library 90

2.3 Ligand Based Virtual Screening Techniques 92

2.3.1 Fingerprint Similarity Searching 93

2.3.1.1 Molecular Fingerprints 94

2.3.1.1.1 ECFP 94

2.3.1.1.2 FCFP 98

2.3.1.1.3 MACCS 99

2.3.2 Turbo Similarity Searching 102

2.3.3 Bioisostere Substructure Searching 105

2.3.4 Principle Component Analysis 116

2.3.5 Naïve Bayesian Classification 127

2.3.6 Decision Tree Analysis 146

2.4 Merging of the Results 157

2.5 References 159

2. Ligand Based Virtual Screening Methods

With the continued development of resistance to existing drug classes, there is an ever increasing need for new antimalarial compounds which act against novel targets. *Pf*bc_1 is one such target, confirmed through the study of acridinediones and other such compounds, which inhibit parasite mitochondrial function in the nM range.[1, 2] ATOV (fig. 1.17) is currently the only licensed *Pf*bc_1 inhibitor, used in combination with proguanil (fig. 1.12) under the branding Malarone.[3-6] Compounds which inhibit *Pf*bc_1 disrupt the ETC by inhibiting the intrinsic membrane protein within the mitochondria, preventing the Q-cycle and thus the generation of ATP, which plays a critical role in respiration. Inhibition of ATP generation therefore results in parasite cell death.[7-9] Of critical importance however is that bc_1 inhibitors remain selective for the parasite and not human bc_1, as this can lead to issues with regard to toxicity. To this end a wide range of LBVS techniques were employed and will be discussed in this chapter. The objective was to identify novel antimalarial chemotypes active against *Pf*bc_1. These methods were then combined as part of a consensus study, as detailed in Chapter III.

2.1 Identification of Initial Data

The virtual screening of chemical libraries has become an essential tool for identifying lead compounds.[10-14] As stated in Chapter I, virtual screening utilises the similarity principle, which states that structurally similar compounds are more likely to exhibit similar properties.[15-18] Through many successful applications, virtual screening has proven to be a rapid and cost effective strategy for evaluating large virtual databases of chemical compounds.[19, 20]

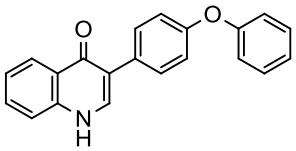
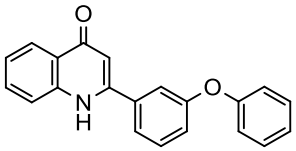
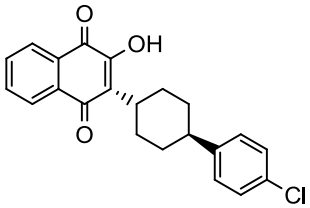
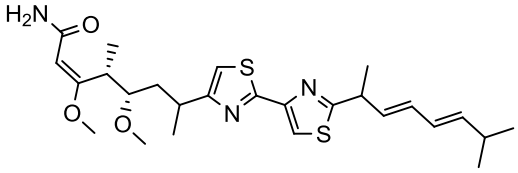
Antimalarial drug discovery efforts within the Chemistry Department at The University of Liverpool have predominantly been concerned with the synthesis and testing of compounds around a particular chemotype, to enable a SAR study and subsequent optimisation around the chemotype. Through collaboration with the Liverpool School of Tropical Medicine (LSTM), compounds can be screened against a required assay, the results from which then drive forward further investigation. There therefore exists vast quantities of biological and chemical data waiting to be utilised, and it is this information which can form the ideal starting point of a LBVS study.

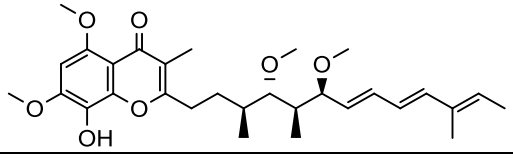
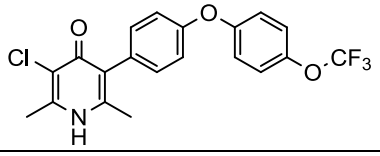
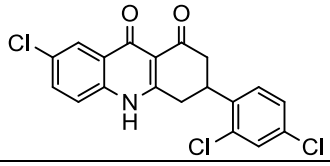
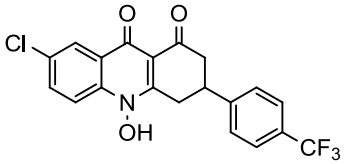
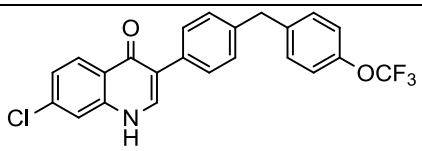
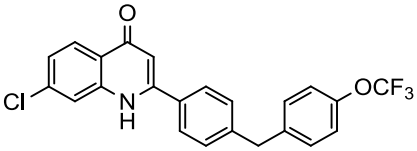
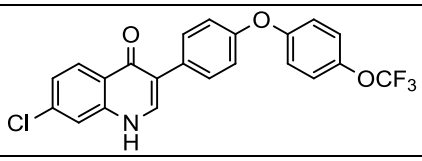
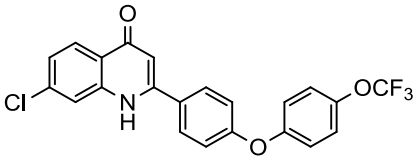
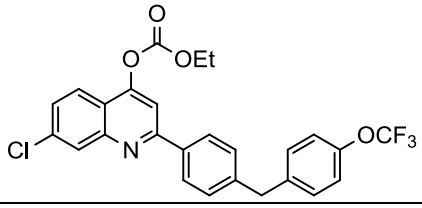
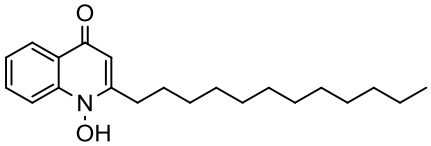
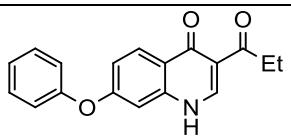
With the biochemical target of interest known (*Pf*bc$_1$), it was necessary to identify suitable compounds for use in virtual screening. At LSTM, compounds are routinely screened for their whole cell growth inhibition, that is, their reported inhibition of the 3D7 CQS parasite. Given that the objective was to find compounds which selectively inhibit *Pf*bc$_1$, the data for virtual screening needed to have been tested against this assay. Unfortunately however, owing to the difficulties, expense, and time required to obtain sufficient amounts of purified *Pf*bc$_1$, compounds are not routinely screened against this particular bioassay, and therefore accurate quantitative IC_{50} values against *Pf*bc$_1$ were only available for a limited number of compounds.

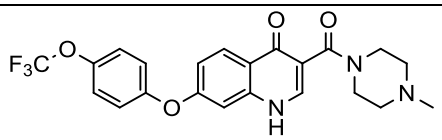
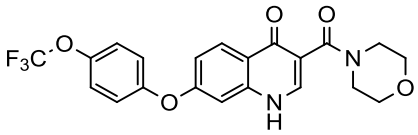
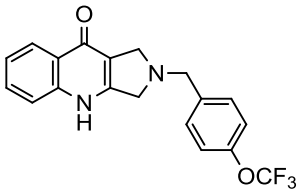
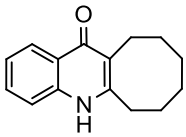
The compounds which had been tested against *Pf*bc$_1$, and therefore used during virtual screening are reported in table 2.1, together with their quantitative *Pf*bc$_1$ activity values and qualitative classifications.[21] These compounds varied significantly in structure and activity, with some active in the single nM range (i.e. Freddie-2-aryl), with others up to several μM (i.e. BC029). Whilst compounds such

as ATOV and GW844520 are known inhibitors of *Pf*bc_1, most of the others were novel structures synthesised at Liverpool. Given that the compounds varied in their activity values so widely, it was necessary to place qualitative cut offs with which to define compounds as either active or inactive. Whilst it is true that all of the compounds exhibited some *Pf*bc_1 inhibition, in order to maximise the enrichment of virtual screening, only the most potent compounds were defined as active. Several of the virtual screening methods employed also required qualitative results, rather than quantitative. Ultimately, compounds were considered active if they had *Pf*bc_1 IC_{50} values of less than 100 nM, or if they exhibited complete parasite bc_1 inhibition in the nM range. Those with IC_{50} values greater than 100 nM, or which exhibited μM inhibition were considered inactive. Using these constraints, of the nineteen compounds tested against *Pf*bc_1, twelve were active and the other seven inactive.

Table. 2.1 Compounds with known activity against *Pf*bc_1.

Name	Compound	*Pf*bc_1 IC_{50} Activity	Classification
Freddie-2-aryl		40.9 nM	Active
Freddie-3-aryl		52.6 nM	Active
Atovaquone		2.7 nM	Active
Myxothiazol		3.5 ± 0.5 nM	Active

Stigmatellin		12 ± 1 nM	Active
GW844520		32 ± 13 nM	Active
WR249685 (S enantiomer)		3 ± 2 nM	Active
Floxacrine (racemic)		802 ± 183 nM	Inactive
Ruan 1		60% inhibition at 1.4 µM	Inactive
Ruan 2		3.5 nM	Active
Ruan 4		32% inhibition at 1.4 µM	Inactive
Ruan 10		62% inhibition at 1.4 µM	Inactive
Ruan 11		60% inhibition at 1.4 µM	Inactive
HDQ		25 nM	Active
DRUG 1		Complete inhibition at 2.8 µM	Inactive

DRUG 2		Complete inhibition at 281 nM	Active
DRUG 3		Complete inhibition at 281 nM	Active
DRUG 6		Complete inhibition at 281 nM	Active
BC029		9.28 µM	Inactive

2.2 Chemical Library

A chemical library is a collection of compounds readily available for use in HTS and virtual screening,[22] and can be widely used for the exploration of chemical space.[23, 24] As stated in Chapter I, chemical space is the space spanned by all possible molecules, with the total number of possible small drug like molecules that populate chemical space estimated to exceed 10^{60}.[25] With such a vast amount of space it is understandable that its exploration has been limited, with only around 60 million small molecules registered with the Chemical Abstracts Service (CAS) as of September 2011.[26] Though the systematic exploration of chemical space is possible through *in silico* databases of virtual compounds,[27, 28] it is expected that much of chemical space contains nothing of biological interest, with searches around specific and focussed areas potentially yielding better results.[29]

Chemical libraries are available from agencies such as the National Cancer Institute (NCI),[30] who have structures for hundreds of thousands of compounds. However,

many other commercial companies also have libraries of varying sizes available. Perhaps one of the most complete, or at least largest chemical library resources is that of ZINC.[31] ZINC is a free database of commercially available compounds ready for virtual screening, compiled from purchasable compounds across numerous sources. In its entirety, it currently consists of over thirteen million compounds whose structures and vendor details, as well as other physicochemical properties are readily available for download and use.

Subsets of the full thirteen million compounds exist that are amenable to particular virtual screening needs. For the LBVS work described here the ZINC lead like library[32] of compounds was chosen, which at the time this work commenced consisted of 2,710,002 unique compounds (version 7). Given that this was the lead like library, a number of filters had previously been applied in order to identify the most lead like compounds from ZINC. Lead structures represent important chemotypes for drug development, ones that are generally pharmacologically active, and consist of simple chemical features amenable to chemical optimisation.[33] By utilising considerations put forward to evaluate drug likeness,[32] filters were applied to the entirety of ZINC to establish this subset of lead like compounds. Molecules passed if their log *P* value was greater than or equal to 2.5, but no more than 3.5, whilst their MW was greater than or equal to 250, but not more than 350 Da. These filters were used as the optimisation of low potency leads is often accompanied by an increase in MW and lipophilicity as a consequence of affinity enhancement, thus making μM hits suitable for optimisation.[32] Additionally, there are no fewer than 5 and no more than 7 rotatable bonds (RBs) in any of the molecules contained in the lead like library, which is in line with Veber's guidelines for oral bioavailability.[34, 35]

Veber found that reduced molecular flexibility, as measured by the number of RBs, was an important consideration for good oral bioavailability, as was a low polar surface area (PSA). The observations were that compounds which have ten or fewer RBs and a PSA of equal to or less than 140 $Å^2$, have a higher probability of good oral bioavailability. Reduced PSA was also found to correlate better with an increased permeation rate than log *P*, with an increase in the RB count having a negative effect on the permeation rate. It has also been found that *in vitro* ligand affinity decreases 0.5 kcal/mol on average for every two RBs.[34, 36]

2.3 Ligand Based Virtual Screening Techniques

With a suitable chemical library selected and a number of seed molecules identified, various LBVS approaches could now be applied, and given that there is currently no crystal structure of *Pf*bc_1 available, this was the most appropriate direction with which to conduct research to find novel antimalarial chemotypes. With the structural information for a number of compounds tested against *Pf*bc_1 known, a whole host of ligand based and classification techniques were available for use. In combination, these techniques were used in a consensus approach, allowing emphasis to be placed onto particular molecules which were selected from multiple screening methods. It involved combining highly disparate properties in order to improve performance by enriching the data.[37] In total, six methods were used and performed in parallel against the ZINC lead like library, based on the active and inactive query compounds (table 2.1). These methods together with the results will now be discussed.

2.3.1 Fingerprint Similarity Searching

Fingerprint similarity searching relies heavily on the similarity principle, which states that structurally similar compounds are more likely to exhibit similar properties.[15-18] However, it is this principle which has been exploited throughout most of the LBVS work. Fingerprint similarity searching is perhaps the simplest of the LBVS methods used, requiring only the identity of the active compounds (table 2.1) to use as queries, and the ZINC lead like library to screen. Fingerprint similarity searching was performed using 2D fingerprints,[14] with similarity between the queries and those in the library quantitatively assessed using the Tanimoto coefficient.[38, 39]

Pipeline Pilot Student Edition v6.1[40] was used to develop a protocol with which to perform fingerprint similarity searching, as it allows for the creation of workflows for the processing of data, including chemical data. The chemical structures of the twelve active compounds were first drawn with ChemBioDraw,[41] and then exported as SMILES (Simplified Molecular Input Line Entry System).[42, 43] SMILES is a form of linear notation used for describing chemical structures of molecules. The ZINC lead like library was similarly manipulated in its SMILES format (readily downloadable), which was particularly useful given its size, as SMILES can encode for many structures yet requiring little storage space. Within the workflow the active molecules were tagged as reference structures, and it was these reference structures which were then used to screen the ZINC lead like library for similar compounds. Molecular similarity between the reference structures and those in ZINC was assessed using molecular fingerprint.

2.3.1.1 Molecular Fingerprints

Molecular fingerprints[44] are representations of chemical structures, originally designed to assist in chemical database substructure searching.[38, 45] Molecular fingerprints have since found use in similarity searching,[46] clustering,[47] and classification.[48] Several methods are available, including Extended Connectivity Fingerprints (ECFP), Functional Class Fingerprints (FCFP),[49] and MDL Public Keys, which have been shown to be effective in similarity searching applications.[50, 51]

2.3.1.1.1 ECFP

ECFPs are a recently developed fingerprint methodology, explicitly designed to capture molecular features relevant to molecular activity.[49] They were first introduced in 2000 with the introduction of Pipeline Pilot,[52, 53] and have since been applied to a broad range of scientifically relevant problems using a variety of analysis methods. Whilst not designed for substructure searching, they are well suited to tasks related to predicting and gaining insight into drug activity,[54] and can be used much like other fingerprint methods for similarity searching, clustering and virtual screening.

ECFPs are derived using a variant of the Morgan algorithm,[55] which was proposed as a method for solving the molecular isomorphism problem. In the Morgan algorithm, an iterative process assigns numeric identifiers to each atom, at first using a rule that encodes the numbering invariant atom information into an initial atom identifier, and later using the identifiers from the previous iterations. Thus, identifiers generated are independent of the original numbering of the atoms, with

the iterative process continued until every atom identifier is unique, or at least as close to unique as symmetry allows. The intermediate results are discarded and the final identifiers provide a canonical numbering scheme for the atoms. Initially, each atom is assigned a connectivity value equal to the number of connected atoms.[14] In the second and subsequent iterations, a new connectivity value is calculated as the sum of the connectivity values of the neighbours. The procedure is repeated until the number of different connectivity values reaches a maximum. In the case of aspirin (fig. 2.1), there are initially three different connectivity values (1, 2, 3). This increases in the subsequent iterations to six, eight and finally eleven. The atom with the highest connectivity value is then chosen as the first atom in the connection table, with its neighbours then listed in order of their connectivity values, and then their neighbours listed and so on. If a tie occurs then additional properties such as atomic number and bond order are considered. For example, the two oxygen's of the carboxylic acid in aspirin have the same connectivity value, as do the terminal methyl and the carbonyl oxygen of the acetyl group. These conflicts can be resolved by consideration of the bond order and atomic number respectively.

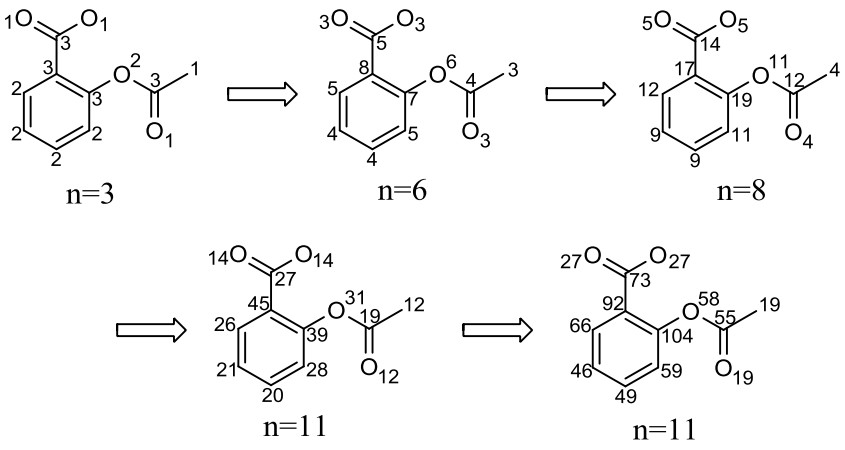


Fig. 2.1 Illustration of how the Morgan algorithm iteratively constructs atomic connectivity values for aspirin. (A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.)

The ECFP algorithm makes several changes to the standard Morgan algorithm. Firstly, ECFP generation terminates after a predetermined number of iterations, as oppose to continuing until identifier uniqueness is achieved. The initial atom identifiers, and all identifiers after each iteration, are collected into a set, and it is this set which defines the extended connectivity fingerprint. Rather than discarding the intermediate atom identifiers, the ECFP algorithm retains them, with finding these partially disambiguated atom identifiers being the goal of the process. This means that the iteration process does not have to proceed to completion, but is performed for a predetermined number of iterations. Secondly, since perfectly accurate disambiguation is not required, algorithmic optimisations are possible. For example, in the standard Morgan process, the identifiers must be carefully recorded after each iteration to avoid mathematical overflow and possible collision of atom environments given the same identifier. However, this recoding has the side effect of creating identifiers that are not comparable between different molecules, whereas in the ECFP algorithm, this computationally expensive step is replaced with a fast hashing scheme. The result of this is saving computational effort for fingerprint generation, and more importantly, allowing for the generation of identifiers which are comparable across molecules.

There are three sequential stages in ECFP generation:[49]

1. An initial assignment stage in which each atom has an integer identifier assigned to it.

2. An iterative updating stage in which each atom identifier is updated to reflect the identifiers of each atoms neighbours, including identification of whether it is a structural duplicate of other features.

3. A duplicate identifier removal stage in which multiple occurrences of the same feature are reduced to a single representative in the final feature list. (The occurrence count may be retained if one requires a set of counts rather than a standard binary fingerprint.)

Atoms are first assigned an integer identifier such as atomic number, and then collected together into an initial fingerprint set. Next, each atom collects its own identifier and the identifiers of its neighbouring atoms into an array, with a hash function applied to reduce this array back into a new, single integer identifier. These new identifiers replace the old and are added into the fingerprint set. This iteration is repeated a predetermined number of times, and once complete, duplicate identifiers in the set are removed and the remaining integer identifiers in the fingerprint set define the ECFP fingerprint.

ECFPs are intended to capture precise atom environment substructural features, with the initial atom identifier for the standard ECFP fingerprint using atom information from the Daylight atomic invariants rule.[43] The Daylight atomic invariants are six properties of an atom in a molecule that do not depend on initial atom numbering.[56] These properties are: the number of immediate neighbours that are heavy (non-hydrogen) atoms; the valence minus the number of hydrogen's; the atomic number; the atomic mass; the atomic charge; and the number of attached hydrogen's. One additional property is also included, and that is whether the atom is contained in at least one ring. To create an integer identifier from this information, these values are hashed into a single 32-bit integer value, and this value is the initial atom identifier.

ECFPs have adopted the convention of being described as ECFP, but in practicality this is followed by an underscore and a number. The appended number is the

effective diameter of the largest feature, and is equal to twice the number of iterations performed. For example, if three iterations are performed, the largest possible fragment will have a width of six bonds, therefore the fingerprint name will end in six (ECFP_6). Increasing iterations extends the environment of the atom to include higher order neighbours, such that ECFP_2 considers the first order neighbours, with ECFP_4 extending this to the third order neighbours.[14]

2.3.1.1.2 FCFP

There exists a variant to the standard ECFP algorithm, termed FCFP. FCFP intends to capture more abstract role based substructural features, derived from the functional class, or pharmacophore role of the atoms in a molecule. The highly specific atom information contained in the initial atom identifiers for ECFPs, allows the generation process to rapidly discover identifiers that represent a broad set of precisely defined structural features. However, for some purposes, this specificity may be undesirable, and some level of abstraction useful. For example, a chlorine or a bromine substituent on a ring may be functionally equivalent but would be distinguished by the ECFP process. It may therefore be preferable to have all halogens appear as equivalent atom types in the fingerprint process, and similarly for all hydrogen bond acceptors or donors to appear equivalent.

FCFPs are generated using a more abstract, pharmacophoric set of initial atom identifiers.[57] Each atom is identified by a six-bit code, where a given bit is considered "on" if the atom plays the associated role. These atom roles are: hydrogen bond donor or acceptor; negatively or positively ionisable; aromatic; halogen. An atom could potentially have more than one role, or none at all. The six-

bit code for each atom is the initial atom identifier, and once these are calculated the process of calculating FCFPs proceeds identical to that of the ECFP process. Similarly, FCFPs are also followed by a number indicating the number of iterations performed.

2.3.1.1.3 MACCS

Developments in molecular database optimisation led to the first explicitly binary fingerprint for substructure searching, used by the MACCS system from Molecular Design Limited (MDL).[58] Molecular descriptors can be encoded into binary keybits, with either a one-to-one relationship between descriptors and keybits, or using hashing to create a many-to-one or many-to-many relationship. An ordered collection of keybits constitutes a keyset, which have been used successfully in substructure searching.[45] For example, a number of topological features in a query molecule can be used to set keybits, with a search for molecules matching that query being used to screen out all molecules in the database which do not set those same keybits. Other keybits include the assessment of an atom to see whether it is a halogen, and they may even comment on the nature of its neighbours. Two particular MDL keysets have been widely explored, one containing 960 keybits, the other 166 keybits.[51, 58] The keyset containing 166 bits is a subset of the full 960 keys,[59] and is called the MDL public keys. It was developed for the rapid substructure searching of databases, and are readily calculated in Pipeline Pilot.[40] These keysets have found use in a variety of drug discovery techniques including QSAR[59] and virtual screening,[60] but the design of new keysets continues to be of interest, with the examination of *in vitro* affinity fingerprints,[61, 62] *in silico* affinity

fingerprints,[63, 64] and feature trees[65] as methods of producing keysets optimised for similarity searches.

Within Pipeline Pilot Student Edition v6.1[40] there are several molecular fingerprint methods available. Even the most intelligently selected set of keys is limited in its coverage.[51] To this end the fingerprint similarity searching of the reference compounds in the ZINC lead like library was repeated to incorporate the results from multiple fingerprint methods. In total, seven fingerprints were used: ECFP_2; ECFP_4; ECFP_6; FCFP_2; FCFP_4; FCFP_6; MDLPublicKeys. The molecular fingerprints simply compare each of the reference structures to the compounds in the chemical library. To identify those which were most similar a cut-off parameter was required. To assess the level of similarity the Tanimoto coefficient (Chapter I) was used, as this has been successfully applied across many virtual screening methods.[66] The cut-off parameter was as follows: to only consider a compound from ZINC sufficiently similar if it had a minimum Tanimoto coefficient value of 0.7, and a maximum value of 0.99, compared to one of the reference structures. By searching for novel chemotypes we are looking for molecules to retain some similarity to one of the known actives (utilising the similarity principle), but that they not be identical (hence the maximum value of 0.99). Virtual screening aims to move through chemical space in an informed manner, using existing knowledge to make rational decisions. By selecting a minimum Tanimoto coefficient value of 0.7 (that is 70% of the binary fingerprints are set to "1"), this would allow for some similarity between the molecules, but also incorporating diversity into their structures. Diversity will also arise given the way in which the seven different molecular fingerprint methods

are calculated. This cut-off is also supported across the literature.[67-72] The number of hits from ZINC for the different fingerprint methods is detailed in table 2.2.

Table. 2.2 Results of fingerprint similarity searching.

Molecular Fingerprint Method	Number of Hits from ZINC
ECFP_2	21
ECFP_4	1
ECFP_6	0
FCFP_2	1161
FCFP_4	17
FCFP_6	3
MDLPublicKeys	11133

It was found that increasing the number of iterations for a particular fingerprint method did not identify any new molecules (i.e. all of the hits from ECFP_4 had already been identified by ECFP_2). By extending the number of iterations, fingerprints became more specific and thus, when screening a chemical library, fewer compounds were considered similar due to these extended and precise fingerprints. As the number of iterations increased, the number of newly discovered identifiers decreased, until eventually no new molecules were discovered, as observed with ECFP_6. It has previously been suggested that fewer than two iterations is sufficient for similarity searching.[49] Due to this, only the results from ECFP_2, FCFP_2 and MDLPublicKeys were considered as these all identified unique sets of compounds, with further iterations giving no new chemical information. Recent research has however shown that these fingerprint methods do indeed explore different areas of chemical space, with each contributing its own unique information.[73]

The results from these methods were merged in Pipeline Pilot Student Edition v6.1,[40] and structural duplicates removed such that only 11,655 unique compounds were identified from fingerprint similarity searching. These unique compounds went on

to form part of the consensus analysis to identify those of most interest across the LBVS work. Fingerprint similarity searching was performed according to the *'Fingerprint Similarity Searching Protocol'* as detailed in the Experimental Chapter.

2.3.2 Turbo Similarity Searching

It is understood that the effectiveness of similarity methods can vary greatly from problem to problem, so the use of several different methods is recommended when performing similarity searches.[74-76] Turbo similarity searching can be seen as an expansion of fingerprint similarity searching, and may be useful when details of just a single active compound are available.[77, 78] It can be useful in exploring the chemical space available, provided that the query compounds are not too tightly clustered, and incorporate an element of structural diversity. Turbo similarity searching uses information about the nearest neighbours to an active structure in a conventional similarity search, to increase the effectiveness of virtual screening.[79]

Fingerprint similarity searching is used to initially identify the nearest neighbours of a query structure to those in a chemical library. These nearest neighbours are then used as the seeds in a new similarity search, with the resulting hit lists being combined. In its simplest form, turbo similarity searching involves moving through chemical space by considering the nearest neighbours, of the nearest neighbours. The rationale behind turbo similarity searching is once again based upon the similarity principle, as it is assumed that the nearest neighbours of the initial query compound are themselves active, and as such may also be used as queries for subsequent similarity searching.[14] Studies have shown that consistent improvement in performance can be achieved over more conventional similarity searching

methods based on the initial search only, and that multiple queries, such as the presumption that nearest neighbours are active, can be combined to increase the enrichment.[80] Figure 2.2 gives a simple representation of how turbo similarity searching works, with the active molecule used as the initial query to identify the nearest neighbours, and then these themselves used to identify additional nearest neighbours.

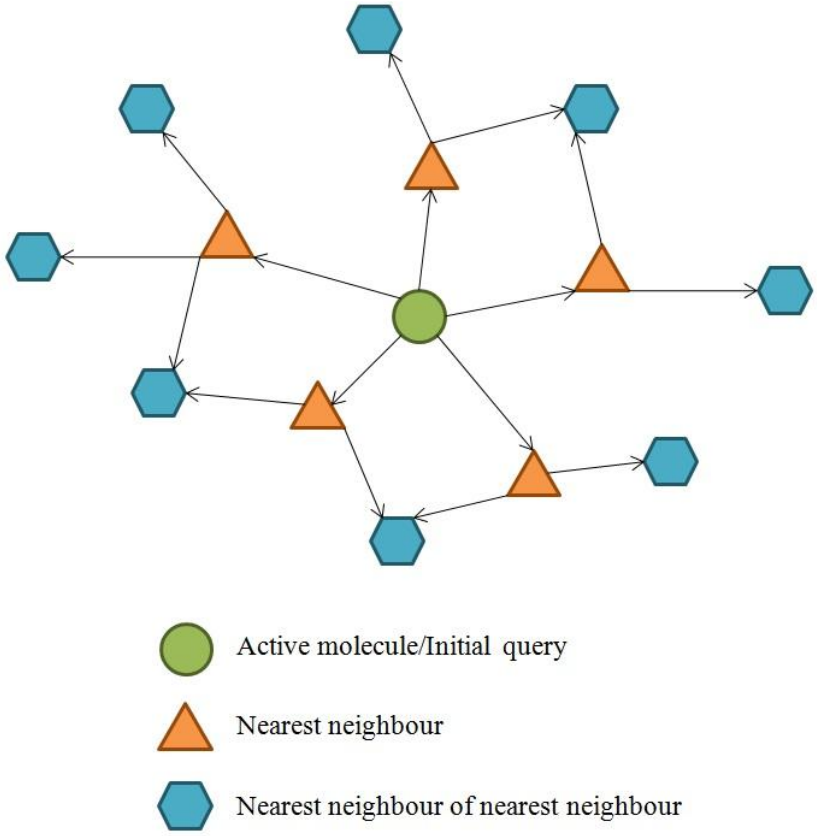


Fig. 2.2 Illustration of turbo similarity searching.

Pipeline Pilot Student Edition v6.1[40] was used to develop a suitable protocol for turbo similarity searching, with the twelve active compounds (table 2.1) tagged as reference structures, and the ZINC lead like library used for screening. Molecular similarity between the two was assessed using the ECFP_2, FCFP_2 and MDLPublicKeys fingerprints. The Tanimoto coefficient was used to assess the similarity between the reference structures and the hits from ZINC; however, a more

stringent cut-off parameter was put in place. Whereas with fingerprint similarity searching molecules were required to have a Tanimoto coefficient value greater than only 0.7, for turbo similarity searching this was increased to 0.8. This was done to fall in line with the very nature of turbo similarity searching, as the method is concerned with moving further out into chemical space, and is heavily based upon the assumption that the nearest neighbours of the active query may potentially be active. This is an optimistic assumption, as it would be highly unlikely that every hit from fingerprint similarity searching would yield an active result, thus the value was raised. A cut-off of 0.8 will still allow for some level of diversity in the reported structures, and indeed expand upon the chemical space being sampled, but without straying too far from the initial, active query, thus hopefully retaining some of the chemical features which made that structure active.

Besides the alteration in the lower limit of the Tanimoto coefficient parameter, the protocol was initially identical to that of the fingerprint similarity search. Where it differed was that instead of terminating after one search, a further iteration was performed, with the nearest neighbours of the initial query (the hits), used to perform another similarity search of the chemical library using the same fingerprint method. Similarity was assessed using the Tanimoto coefficient. To control the amount of space being searched, a cap was placed on the number of nearest neighbours to be used as queries in the second iterations of similarity searching. Only the top 250 scoring compounds from the initial search (those closest to an active structure according to their Tanimoto coefficient) were used as seeds for an additional search. With regard to ECFP_2 and FCFP_2, this did not pose too much of a concern as the first search only produced 3 and 76 hits respectively. However, when using

MDLPublicKeys the cap had to be employed as there were 336 hits. When both rounds of similarity searching were complete, the hits were merged and duplicate molecules removed. The results for these searches are shown in table 2.3, with the number of hits from ZINC reported for each fingerprint method.

Table. 2.3 Results of turbo similarity searching.

Molecular Fingerprint Method	Number of Hits from ZINC
ECFP_2	19
FCFP_2	1053
MDLPublicKeys	13261

The number of hits for each method is comparable to those from fingerprint similarity searching (table 2.2). However, had the minimum Tanimoto value been left at 0.7, then the number of hits for ECFP_2, FCFP_2 and MDLPublicKeys would have been 204, 44,869 and 150,716 respectively, greater than 10 fold the number of hits in all cases. So, whilst the reasoning behind the higher cut-off was sound, it had the additional benefit of reducing the number of hits, thereby reducing the potential noise in the data. As before the compounds from each of the different fingerprint methods were merged and duplicate molecules removed, resulting in 13,771 unique compounds identified by this method, ready for inclusion in the consensus analysis. Turbo similarity searching was performed according to the *'Turbo Similarity Searching Protocol'* as detailed in the Experimental Chapter.

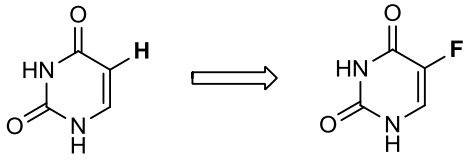
2.3.3 Bioisostere Substructure Searching

Isosteres are atoms or groups of atoms that have the same valency, and which have similar chemical or physical properties.[81] Examples include silicon and carbon, and carbon dioxide (CO_2) and nitrous oxide (NO_2). More specific examples in terms of functional groups include the understand that SH, NH_2 and CH_3 are isosteres of OH,

and that S, NH, and CH_2 are isosteres of O.[82] Isosteres can be used to determine whether a particular group is an important binding group or not, by altering the character of the molecule in as controlled a way as possible. For example, replacing O with CH_2 will make little difference to the size of the analogue, but may have a marked effect on its polarity, electronic distribution and bonding. Whilst replacing OH with the larger SH may not have such an influence on the electronic character, but steric factors become more significant. Isosteric groups may also be used to determine whether a particular group is involved in hydrogen bonding, such that replacing an OH with CH_3 would completely eliminate hydrogen bonding, yet a NH_2 group would not.

A biologically active compound containing an isostere is called a bioisostere. These are frequently used in drug design as they may still be recognised and accepted by the body, but its function will be altered compared to the parent molecule. This is through varying the character of a molecule in a rational way, with respect to features such as size, polarity, electronic distribution and bonding.[83, 84] They have also been employed by chemists for the modification of lead properties. A bioisostere is a group that can be used to replace another group whilst retaining the desired biological activity. They are often used to replace a functional group that is important for target binding but is problematic in another way. A good example of successful and simple bioisostere replacement is that of the anti-tumour drug 5-fluorouracil (fig. 2.3).[85] The drug is a pyrimidine analogue which acts as an irreversible inhibitor of thymidylate synthase, blocking the synthesis of pyrimidine thymidine which is a nucleotide essential for deoxyribonucleic acid (DNA) replication. 5-flouoruracil therefore causes cancer cell death due to a lack of

thymidine required for replication. Fluorine is often used as an isostere of hydrogen since it is similar in size (Van der Waals radii of 1.47 and 1.20 Å respectively).[86] However, it is more electronegative and can be used to vary the electronic properties of a drug. The presence of fluorine in place of enzymatically labile hydrogen can disrupt the enzymatic reaction, as C-F bonds are not easily broken.[87] It was this understanding which was utilised in the manipulation of uracil, such that 5-fluorouracil is accepted by the target enzyme as it appears almost identical to the natural substrate, but the mechanism of the enzyme-catalysed reaction is totally disrupted as the fluorine has replaced the hydrogen, which is normally lost during the mechanism.

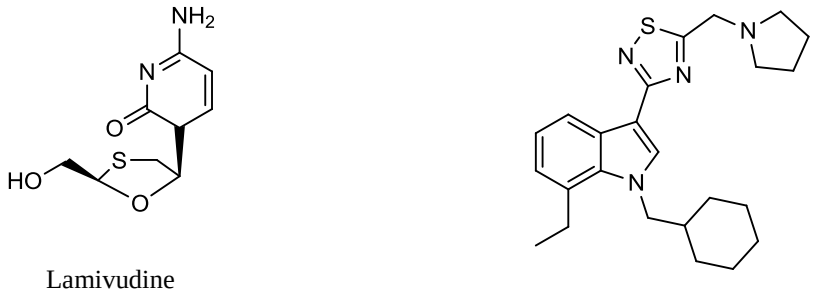


Uracil 5-fluorouracil

Fig. 2.3 Uracil and 5-fluorouracil.

Bioisosteric replacement and scaffold hopping are well used within drug design to improve the synthetic accessibility, potency and drug like properties of compounds, as well as to move into and around novel chemical space.[88] Bioisosteric replacement is concerned with the swapping of functional groups that have similar properties, whereas scaffold hopping involves the replacement of the core framework in a molecule to try and improve the properties of that molecule. Additional examples of the successful use of bioisosteric replacement within drug discovery includes the design of novel nucleoside derivatives in the search for new drugs active against the human immunodeficiency virus (HIV).[89] Derivatives were synthesised based on bioisosteres of lamivudine (fig. 2.4), which is used for the treatment of the hepatitis

B virus (HBV) and HIV. More recent work includes the design of novel indole-3-heterocycles, with the one in figure 2.4 found to be a potent CB1 cannabinoid receptor agonist, a member of the GPCR superfamily.[90] CB1 is a potential therapeutic target against pain, glaucoma, brain injury, multiple sclerosis and obesity,[91] with the novel indole-3-heterocycle in figure 2.4, synthesised based on a bioisosteric replacement of the piperazine amide heterocycle group. This led to improved *in vitro* and *in vivo* stability, as well as an increased duration of action.



Lamivudine

Fig. 2.4 Lamivudine and a novel indole-3-heterocycle.

Owing to its successes within drug discovery, bioisostere substructure searching was deemed of potential interest with regard to this research and was performed based on the quinolone core substructure detailed in figure 2.5. Of the twelve active compounds, eight of them contained this quinolone core structure, and thus was the most common fragment/chemotype amongst the actives, making it the most appropriate query. The groups labelled **A** in figure 2.5 represent possible attachment points which were taken into consideration in the template query.

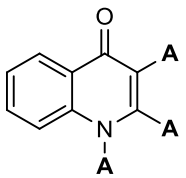


Fig. 2.5 Quinolone core template, where **A** represents possible attachment points.

BROOD 1.1.2[92] was used to identify bioisosteres of the query structure by searching fragment databases using a number of different metrics of similarity. Fragment databases may originate from the deconstruction of molecular databases using pseudoretrosynthesis,[93] and can be used in bioisosteric *de novo* design through comparison of parts of the active lead structure, to fragments abstracted from pharmacologically active molecules. By creating fragment libraries from databases of biologically active molecules, this allows for the identification of building block fragments that are rich in biologically recognised elements and privileged motifs and structures.[93] Fragment based screening inherently incorporates chemical variety, and has been found to decrease the drug discovery process timescale, and even improve the understanding of the pharmacophore at the active site.[94]

BROOD[92] pioneered the technique of fragment replacement from the perspective of shape based template comparison. It was designed to help explore chemical and property space around a lead structure, generating analogues of the lead by replacing selected fragments in the molecules with fragments that have similar shape and electrostatics. The software allows a user to enter a single query fragment and search a large database of known molecular fragments in order to identify fragments that are similar. Each database fragment is compared to the query fragment in 3D with regard to its shape, chemistry, electrostatics and geometric presentation of attachment groups. The best fragments in each of the four classes are then saved as potential bioisosteres. Four searching methods were used to search for bioisosteres of quinolone, each encompassing different structural criteria as follows:

1. color – Those with the best overlap of shape, atom-types and attachment geometry.

2. elect – Those with the best overlap of shape, electrostatics and attachment geometry.

3. struc – Those with the best replication of the attachment geometry (ignoring chemical properties).

4. queryAnalog – Those that are close analogues of the query fragment.

The results from each search are referred to as hitlists. The color bioisostere hitlist is the most general and practical set of bioisosteres. These fragments are those deemed most similar to the query structure, based on the sum of shape similarity and atom-type similarity, including attachment points. By default, the color force field treats all color interactions with equal weighting. These weights can be used to distinguish between better bioisosteres. The elect or electrostatic bioisostere hitlist is analogous to the color method in its utility. There are however, two critical differences in how the similarity of fragments is evaluated. The first difference is that rather than comparing atom-types, the electrostatic potential projected into the area around the molecules is compared. This comparison is evaluated by the electrostatic overlap between the query structure and the fragments, as determined by the Poisson-Boltzmann equation which describes electrostatic interactions.[95] The second difference is that there is no consideration of the attachment points in the electrostatic term, so an additional term taking the difference in attachment point position into account is added. However, the elect calculation is more computationally expensive than that of color. The struc or structure bioisostere hitlist is very specialised, and is concerned with giving the best alignment of attachment points as measured by the root mean square deviation (RMSD) between the query structure and fragments, regardless of shape, chemistry and electrostatics.

The hitlist will report fragments that can present substituents in a similar manner to the query structure. The final bioisostere hitlist is that of queryAnalog. This hitlist contains all fragments that have similar graphs to the query structure. Similar graphs are considered to be all those fragments that have identical uncoloured graphs or uncoloured graphs that differ by only one or two atoms from the query. Each of these fragments is orientated and evaluated by the color method to allow appropriate visual comparison of the fragment to the query. While this hitlist often contains many fragments that are obvious replacements for a query, such as those which would be selected by an experienced medicinal chemist, they may also include fragments with dramatically different chemistry from that seen in the query. BROOD[92] also considers bioisosteric fragments that contain ring systems, so that in practice, this often produces small, rigid, ring systems that quite nicely overlay with linear query fragments.

BROOD 1.1.2[92] contains some 170,000 fragments spread across the f5 (around 140,000 fragments) and f50 (around 30,000 fragments) libraries. The *de novo* generation of fragments may seem attractive but often yields unrealistic chemical fragments.[96] The fragment libraries in BROOD however, are derived from a collection of roughly 12 million commercially available compounds. These contained roughly 1 million unique molecular fragments, each with 15 or fewer heavy atoms, and one to three attachment points. Potentially toxic, reactive or chemically complex fragments were removed, with the remaining half a million fragments ranked according to their frequency in the original 12 million compounds. f50 contains fragments which occurred in 50 or more of the original molecules, with

f5 containing all those fragments which occurred in 5 or more of the original molecules (hence its large size).

Each of these libraries was searched four times for bioisosteres of the quinolone core (fig. 2.5), based on the four different bioisostere searching methods available. Table 2.4 illustrates the results of quinolone bioisostere searching across the two fragment libraries for the four methods. By default, the maximum number of fragments per query for each of the search methods was set to 200. These were ranked by shape and chemistry for their similarity based on the quinolone query. That is, fragments that were more similar to the query were reported first, ensuring only the most appropriate bioisosteres were used. Across the four bioisostere searching methods applied to the f5 library, there were a total of 670 fragments reported. However, when these were merged using Pipeline Pilot[40] there were found to be only 546 unique entries. This is because several of the fragments appeared multiple times (84 appeared twice and 20 appeared three times). Similarly for the f50 library, 619 fragments were identified, but there were again a number of fragments found multiple times, with some 451 unique fragments (96 appeared twice and 36 appear three times).

Table. 2.4 Number of bioisosteres identified from the f5 and f50 fragment libraries for the four different search types.

	Number of Fragments Identified from Library	
Bioisostere Search Method	**f5**	**f50**
color	200	200
elect	200	200
struc	200	200
queryAnalog	70	19
Number of Unique Fragments:	**546**	**451**

When the two unique fragment hitlists were merged the number of unique fragments was reduced further to 848. There therefore appears to be some overlap between the

f5 and f50 fragment libraries. Within Pipeline Pilot Student Edition v6.1[40] these 848 bioisosteres were subjected to analysis to find only those which contained novel fragments. This had two benefits in that it reduced the number of bioisosteres to create a more focused approach, but also removed additional duplicate fragments. The novel fragment types were found using the *'Find Novel Fragments'* component in Pipeline Pilot to identify the unique molecular fragments in the dataset, such that each occurred only once. Novel fragments were defined based on their ring assemblies to find potential drug-like structures amenable to chemical optimisation. 276 novel fragments were identified, with these filtered further to include only fragments with a ring count greater than one. This was to remove simpler heterocyclic structures, and identify unique fragments that looked like potential chemotypes, as inspection of the compounds tested against *Pf*bc$_1$ showed them all to contain fairly rigid ring assemblies. Fragments were also filtered to remove those which contained charged groups, to keep the fragments to be used as bioisosteres as uncomplicated as possible. These filters reduced the fragment set to 166 bioisosteres, several examples of which can be found in figure 2.6 (see appendix for all 166 bioisosteres).

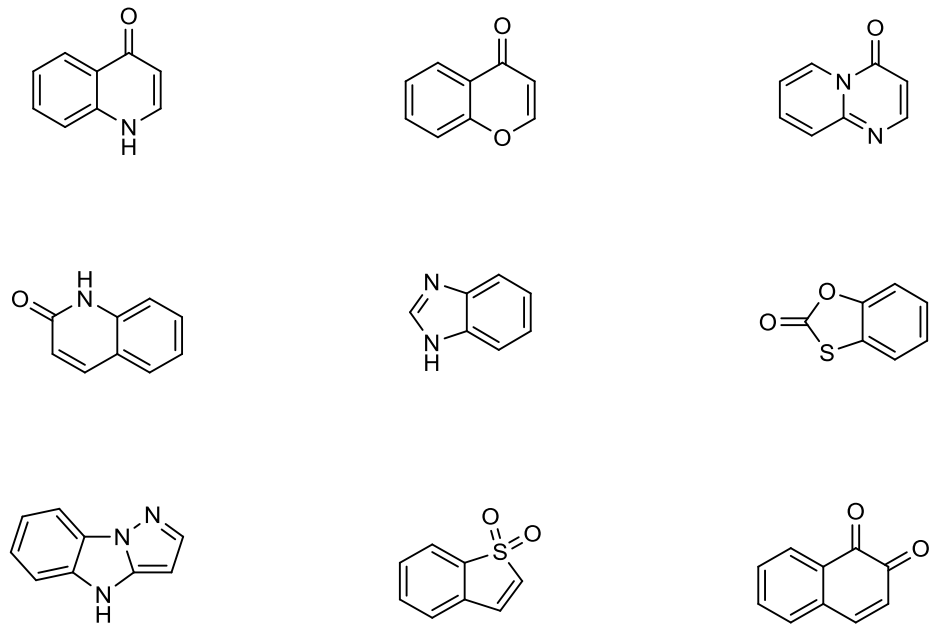


Fig. 2.6 Examples of quinolone bioisosteres identified from the BROOD[92] fragments libraries.

Most of the bioisosteres are simple modifications of the quinolone core structure. However, some of them do vary quite significantly in their appearance, such as the pyrazole containing tricyclic, suggesting that their similarity to quinolone must lie elsewhere, such as in their shape or electrostatic distribution. What is clear though is that the bioisosteres do indeed cover a range of chemical diversity.

With the 166 bioisosteres in hand a substructure search of the ZINC lead like library could be performed. This was done by developing a protocol within Pipeline Pilot Student Edition v6.1,[40] to identify all those molecules within the chemical library that contained one or more of the bioisosteres. Of the 2.7 million molecules in the ZINC lead like library, 466,840 were found to contain a bioisostere, equating to just over 17% of the total number of compounds found in the library. Measures were therefore adopted to reduce this number further, to optimise the diversity across the hits. Of the 166 bioisosteres, 15 did not appear in any of the compounds in ZINC, and as such are not represented. However, of the 151 that did appear the number of

hits per bioisostere varied greatly. 70 of the fragments had between 1 and 200 hits, whilst the other 81 all had greater than 200 hits each. In one instance, a particular bioisostere was found to be present in 46,393 compounds. These results are shown in figure 2.7.

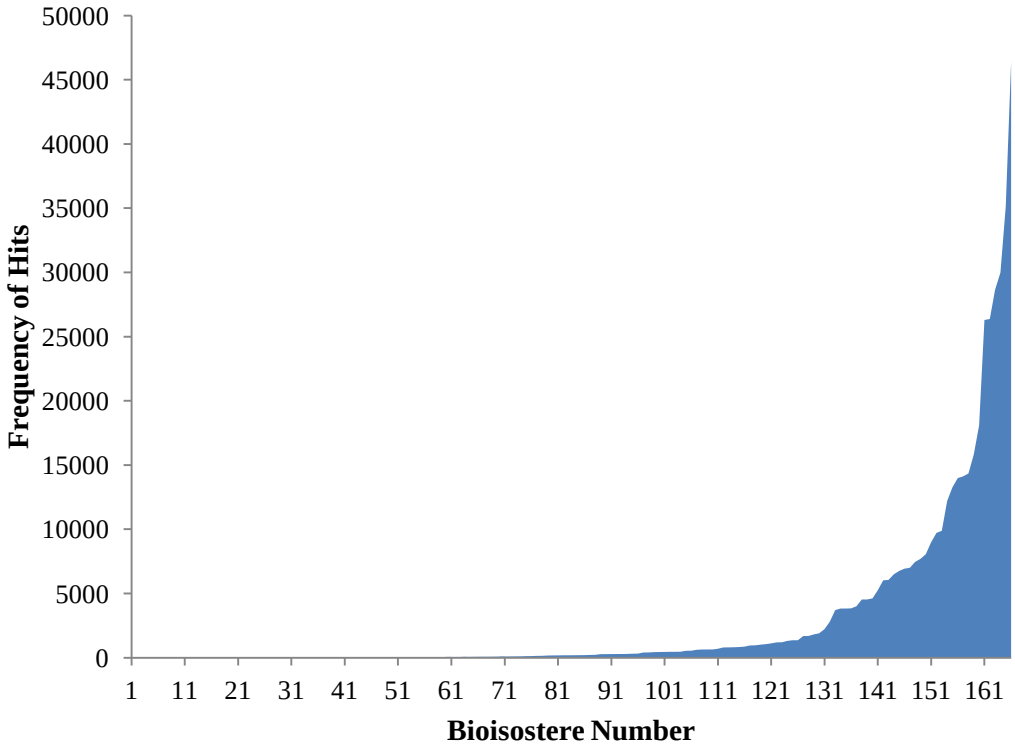


Fig. 2.7 Graph to illustrate the frequency of hits per bioisostere from the ZINC lead like library.

In order to reduce the number of hits to best sample the chemical space, a representative number of molecules was taken when the number of hits exceeded a certain threshold. This threshold was set to 200 hits, so that when the number of hits fell below this threshold for a particular bioisostere, all molecules were taken forward. Yet when the number of hits was greater than 200, a diverse sample of these molecules was taken to get a good representation of the chemical space available. The merits of using a cut-off of 200 could be argued for and against, but it allowed for a relatively similar number of hits to be considered per bioisostere, and

as such, prevented overemphasis of any particular bioisostere. Diversity analysis was performed in Pipeline Pilot Student Edition v6.1,[40] utilising the *'Diverse Molecules'* component by selecting the 200 most diverse molecules from the hits for a particular bioisostere, based on FCFP_4 fingerprints. Following this the hits for each bioisostere were merged and resulted in 20,319 unique molecules. Though slightly larger than the number of hits from fingerprint similarity and turbo similarity searching (11,655 and 13,771 respectively), the number of hits for consideration during consensus analysis is of a similar order of magnitude, and is a much more manageable size when compared with the original 466,840 hits. Also, despite the decreased number of hits, we can still be assured of the chemical diversity described within the dataset owing to the similarity calculations performed. The bioisostere substructure searching was performed according to the *'Bioisostere Substructure Searching Protocol'* as described in the Experimental Chapter.

2.3.4 Principle Component Analysis

The dimensionality of a dataset is the number of variables or molecular descriptors used to describe each of the objects or molecules in that dataset.[14] Principle component analysis (PCA) is commonly used to reduce this dimensionality, especially when there are significant correlations between some or all of the descriptors.[14] PCA was conceived in 1901,[97] and acts by calculating a new set of variables based on the previously calculated molecular descriptors or properties for the dataset. These new variables, termed principle components (PCs), explain much of the variance in the original data, but generally with fewer variables. It is a common statistical technique used for finding patterns in data of high

dimensionality, where graphical analysis would not normally be appropriate. From this it would then be possible to highlight similarities and differences in the data.

The simple 2D example shown in figure 2.8 illustrates the concept of PCs very effectively. As can be seen, the scatter plot shows a high degree of correlation between the x_1 and x_2 values, and it would be possible to explain the variation in this data by introducing a single variable which is a linear combination of these two, such as $z = x_1 - x_2$. This new variable z can be referred to as a principal component.

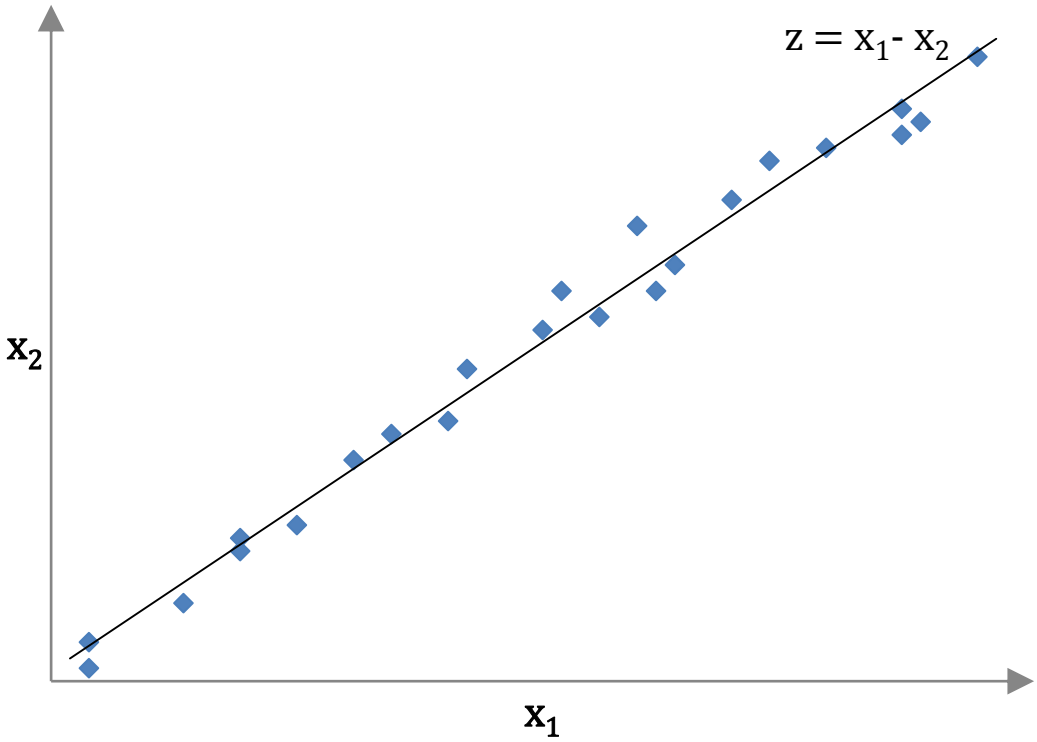


Fig. 2.8 Graph to illustrate principal components in their simplest form.

In the case of a multidimensional dataset, that is one which consists of more than two variables. Each of the PCs comprises a linear combination of the original descriptors as described by equations 2.1.

$$PC_1 = c_{1,1}x_1 + c_{1,2}x_2 + \cdots c_{1,p}x_p$$

$$PC_2 = c_{2,1}x_1 + c_{2,2}x_2 + \cdots c_{2,p}x_p$$

$$PC_i = c_{i,1}x_1 + c_{i,2}x_2 + \cdots c_{i,p}x_p = \sum_{j=1}^{p} c_{i,j}x_j$$

Eq. 2.1 Equations to illustrate the principal components in a multidimensional example.

In equations 2.1, PC_i represents the i^{th} principal component, with $c_{i,j}$ the coefficient of the descriptor x_j, and p being the number of descriptors. The first PC maximises the variance in the data and explains the largest proportion of the variance, thus the data has the greatest spread across the first PC. The second PC accounts for the maximum variance in the data that is not already explained by the first PC, and so on and so forth for any additional PCs. An interesting property of the PCs is that they are all orthogonal to each other, so that the highest possible variance in the data can be explained. With regard to the simple 2D example in figure 2.8, this would mean that the second PC would be uncorrelated to the first, and perpendicular to it.

The total number of PCs which can be calculated for a dataset is equal to the smaller of the number of molecules in the set, or the total number of descriptors. To explain all of the variance in the data it would usually be required to include all of the PCs. However, for most datasets the greatest proportion of variance is explained by only a small number of PCs, due to correlations between the original variables. As the number of PCs increases the additional information explained decreases, until eventually new PCs explain no new variance.

PCs are calculated from the variance-covariance matrix, such that if **A** is the matrix with n rows (corresponding to n molecules) and p columns (for the p descriptors) then the variance-covariance matrix is the $n * n$ matrix $\mathbf{AA^T}$, with superscript $^{\mathbf{T}}$ indicating a transposed matrix in which the rows and columns of the matrix have

been interchanged. The eigenvectors of this matrix are the coefficients $c_{i,j}$ of the PCs, with the first PC which explains the largest amount of variance corresponding to the largest eigenvalues, and the second PC accounting for the second largest eigenvalues and so on. The eigenvalues indicate the proportion of the variance that is explained by each of the PCs, and if the eigenvalues are labelled λ_i, then the first m PCs account for the fraction of the total variation in the dataset as shown in equation 2.2.

$$f = \frac{\sum_{i=1}^{m} \lambda_i}{\sum \lambda_i}$$

Eq. 2.2 Equation to calculate the total variance explained by principal components.

It is usual to autoscale each of the p variables before extracting the PCs (autoscaling will be discussed more fully in chapter VI) so that each variable will contribute a variance of one to the total dataset, which will thus have a total variance of p. This means that if a PC has an eigenvalue less than one, then it can be considered to explain less variance than one of the original descriptors. It is therefore common to only consider PCs with eigenvalues greater than one.

PCA has found many applications in drug discovery. One recent example includes its use in the assessment of chemical diversity.[98] P-glycoprotein is one of the best characterised transporter proteins responsible for the multidrug resistance phenotype exhibited by cancer cells, and as such, represents one of the main barriers to chemotherapeutic treatment in cancer.[99] There is therefore widespread interest in elucidating whether existing drugs may be candidates as P-glycoprotein substrates or inhibitors. Work has been performed concerned with creating a pharmacophore model based on known P-glycoprotein inhibitors. This model was then applied to the DrugBank[100] database. This highlighted a number of molecules of interest,

twelve of which were later found to be novel P-glycoprotein inhibitors. PCA was used to assess the diversity of these hits, and to comment on any potential SARs which existed. Descriptors commonly used in drug like profiling (i.e. log *P*; PSA; MW) were employed, and it was found that three PCs best represented the spread of variance in the data. When a graphical projection of these PCs was analysed, it was shown that there was segregation in 3D space between P-glycoprotein inhibitors and activators, with this observation being utilised to comment on potential areas of chemical optimisation (fig. 2.9).

Fig. 2.9 Example of PCA being used to distinguish between P-glycoprotein inhibitors and activators. Yellow circle: activators; Blue circle: inhibitors. (A. Palmeira, F. Rodrigues, E. Sousa, M. Pinto, M. H. Vasconcelos and M. X. Fernandes, *Chemical Biology & Drug Design*, 2011, **78**, 57-72.)

Earlier work has involved using PCA to analyse a small set of 141 anticancer agents from the NCI, whose mechanisms of action had been well defined.[101, 102] PCA score plots showed distinct clusters of compounds for some of the mechanisms of action examined. More recently, this work has been extended to analyse a much larger database of 25,026 compounds,[103] with PCA proving to be a useful tool in identifying outliers in this dataset, as well as entry errors.[104] This ultimately led to a cleaner and more correct database. It was also found that compounds which had

similar activity profiles against a number of assays were found to group together in PCA space, with dissimilar compounds more scattered across the plot (fig. 2.10). The grouping clearly reflects the different mechanisms of action of the compounds in the dataset.

Fig. 2.10 Example of PCA being used to identify the grouping of anticancer agents with similar activity patterns. (L. M. Shi, Y. Fan, J. K. Lee, M. Waltham, D. T. Andrews, U. Scherf, K. D. Paull and J. N. Weinstein, *Journal of Chemical Information and Computer Sciences*, 1999, **40**, 367-379.)

A more relevant example would be the use of PCA to perform SAR analysis on a number of antimalarial compounds.[105] Eighty structures were found to exhibit varying activities against the malaria parasite, belonging to eight different compound classes, including primary, secondary and tertiary amines, as well as quaternary ammonium and bisammonium salts. The Broto autocorrelation technique was used to calculate a number of molecular descriptors which describe the physicochemical properties of the molecules, such as their molecular geometry, lipophilicity, and ability to give or receive hydrogen bonds.[106] The generated PCA model was found

to explain 98.55% of the variance in the data with only three PCs, and was also found to correctly classify compounds as either poorly active or active with 98% accuracy. Interpretation of the PCA model (fig. 2.11) also illustrated that both the size and shape of the molecules was crucial for antimalarial potency, with bulkier molecules having improved activity. In this respect, axis 1 principally discriminates active (bulkier) from poorly active (smaller) compounds.

Fig. 2.11 Example of PCA being used to determine antimalarial activity. Active compounds ($IC_{50} \leq$ 1.5 μM) are underlined, with the most active ($IC_{50} \leq 0.3$ μM) in italics. (M. Calas, G. Cordina, J. Bompart, M. BenBari, T. Jei, M. L. Ancelin and H. Vial, *Journal of Medicinal Chemistry*, 1997, **40**, 3557-3566.)

Within this thesis, PCA was used to generate PCs for the active compounds based on their physicochemical descriptors. The resulting PCA model was then applied to the ZINC lead like library of compounds to identify potential compounds of interest. Pipeline Pilot Student Edition v6.1[40] was used to generate a protocol with which to perform PCA, based on the structures and properties of the twelve molecules previously described as active against *Pf*bc$_1$ (table 2.1). The following eight

molecular descriptors were calculated within Pipeline Pilot for each of the twelve compounds: AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings. These were chosen as they describe many properties classically considered to describe drug like properties,[107, 108] and whilst several of these descriptors are also well documented for their application within Lipinski's rule of five,[109] recent work has shown that the additional descriptors improve the accuracy of correctly classifying drug like compounds, compared with Lipinski's properties alone.[110] A PCA model was built using these eight molecular descriptors, with requirements defined such that the minimum variance explained by the model be no less than 75%. A successful PCA model was found for the data, the statistics for which are shown in table 2.5. The table shows the variance explained by each PC, as well as the relative contributions from the different molecular descriptors. It was found that only two PCs were required to explain over 75% of the variance in the data (variance explained by two PCs was 81%), with the first three PCs explaining a total variance of 92%. Successive PCs were shown to only marginally improve the explained variance, with the fourth PC explaining only an additional 6%. Therefore the PCA model was comprised of only the first 3 PCs.

Table. 2.5 Details of the principal components from PCA.

Property	PC1	PC2	PC3
StandardDeviation	2.008968248	1.276653102	0.880959517
VarianceExplained	0.576564775	0.232834735	0.110869953
TotalVarianceExplained	0.576564775	0.809399509	0.920269462
ALogP	-6.98E-02	-0.606789183	0.658393792
Molecular_Weight	-0.4232148	0.263800905	0.380984102
Num_H_Acceptors	-0.404689723	0.412420968	-0.116774579
Num_RotatableBonds	-0.436041043	-0.267068747	-0.10482064
Num_Atoms	-0.435556746	0.295048617	0.334732928
Num_Rings	0.384249182	0.40889702	0.237622919
Num_AromaticRings	0.353200539	0.257970423	0.477738711

Figure 2.12 represents the twelve active molecules across 3D PC space. As can be seen, the molecules are well distributed across this region of chemical space, suggesting that they are diverse in nature. There is very little clustering which is encouraging, so it can be deduced that the twelve active compounds are each contributing unique information to the PCA model.

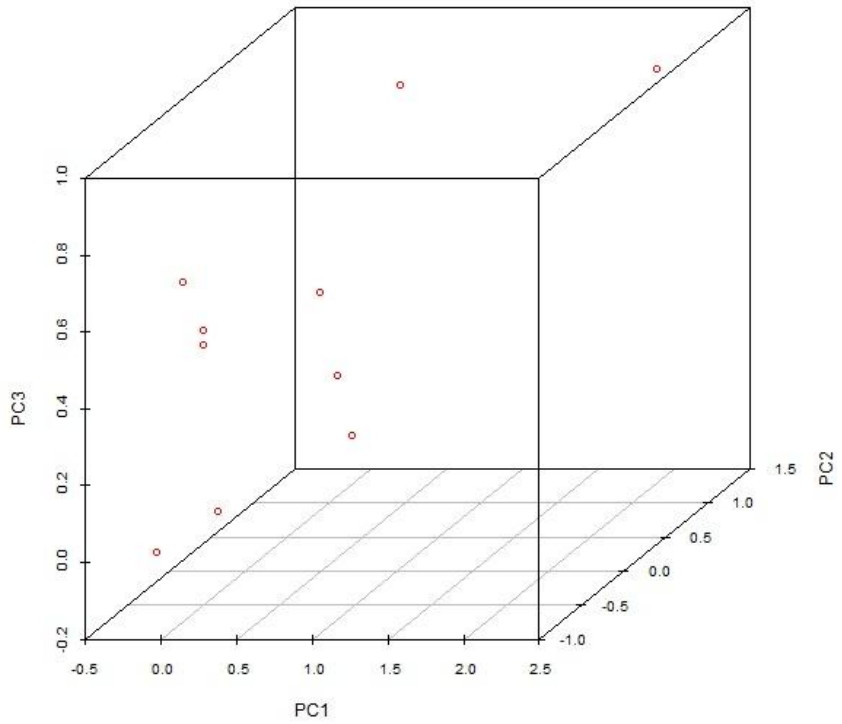


Fig. 2.12 Graph to illustrate the twelve active compounds across 3D principal component space.

The PCA model could next be applied to the ZINC lead like library of compounds to identify any of interest. The same eight molecular descriptors were calculated for each of the molecules in the chemical library, and the PCA model applied to the compounds. With the PC values calculated for the molecules in ZINC they could be plotted as before, to represent the spread of the molecules in 3D PC space. This is illustrated in figure 2.13, and shows the molecules covering a much larger area of PC space, quite closely clustered with a few outliers. However, this is to be expected as we must remember that the structures within the lead like library have already undergone filtering to ensure that only the most promising lead like structures form

part of the chemical library, and therefore already fall under certain physicochemical parameters. Additionally, the distances across the axes in figure 2.13 are much larger than those in figure 1.12, again supporting the conclusion that the compounds in ZINC include a wide array of diverse structures.

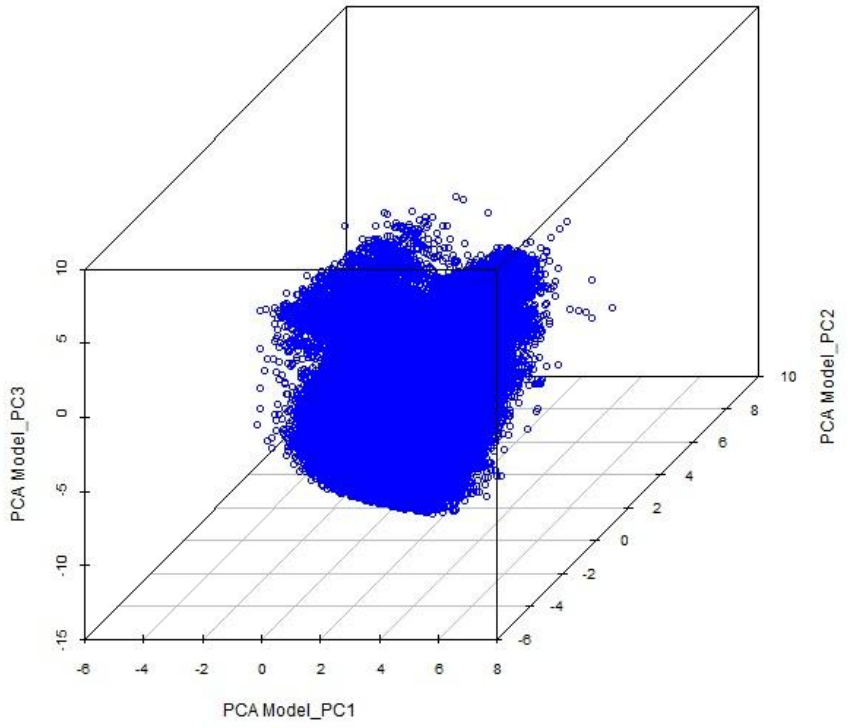


Fig. 2.13 Graph to illustrate the PCA model applied to the 2.7 million molecules of the ZINC lead like library.

With the PCA model applied to the ZINC lead like library of compounds, it was necessary to consider the best way to select compounds of potential interest. To do this the known active compounds were used as reference structures within the 3D PC space. The closest molecules from the ZINC lead like library to these known actives were then selected, with closeness assessed using Euclidean distance. Where the Tanimoto coefficient measures the similarity between two compounds, Euclidean distance essentially measures the distance between two molecules in variable space.[14, 111] In this instance variable space is described by the PCs, and therefore Euclidean distance looks at the distance between molecules in this PC space. Similar

to the Tanimoto coefficient, binary variables can be used to give distance values which range upwards of 0. A Euclidean distance of 0 would indicate that the pair of molecules occupy an identical region of space according to the variables used, but as the value increases they move further away from each other, becoming less close. The Euclidean distance between a pair of molecules using continuous variables can be calculated using equation 2.3, where D_{AB} represents the distance between molecules A and B, and x_i the value of the continuous variables.

$$D_{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2\right]^{1/2}$$

Eq. 2.3 Euclidean distance.

Euclidean distances were assigned to each of the molecules in ZINC, based on which known active molecule it was closest to. In order to try to achieve the correct balance of similarity and diversity, only the 5,000 compounds from ZINC closest to a known active were selected. This was to keep the exploration of chemical space as concentrated and focused as possible. When the PCs of these 5,000 closest compounds was plotted in 3D space (fig. 2.14), the molecules were found to encompass much physicochemical diversity, thus they made for an interesting contribution towards the LBVS consensus study. Principal component analysis was performed according to the *'Principal Component Analysis Protocol'* as detailed in the Experimental Chapter.

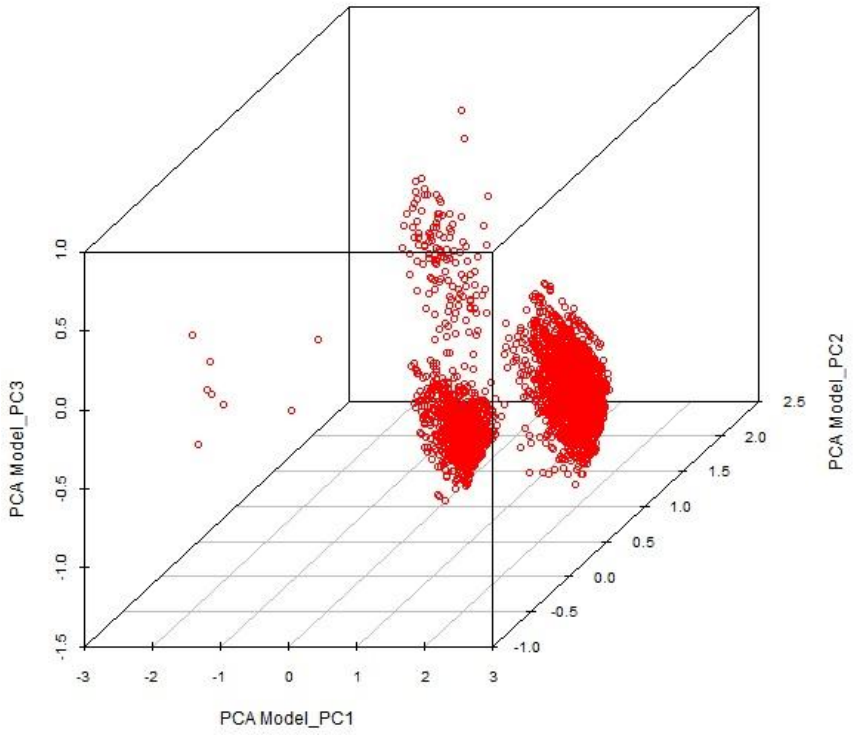


Fig. 2.14 Graph to illustrate the 5,000 closest molecules from ZINC to known active based on their Euclidean distances in principal component space.

2.3.5 Naïve Bayesian Classification

The four LBVS methods described thus far were all concerned with only the structures of the active compounds. The next two methods (Bayesian and decision tree classification) however, utilise the structures and chemical properties of both the active, and the inactive compounds (table 2.1). Whilst individual data points can be of enormous importance with regard to identifying hits (i.e. substructure searching), trends can only be observed if there is sufficient amounts of data available.[112] By using the structural information contained within the active and inactive molecules, trends in the data can be spotted, modelled and exploited, so that each compound in a chemical library can be assessed for its relative merits with regard to the information available. Data mining algorithms such as the application of naïve Bayesian classification have two major aims depending on their particular setting. One is to

help interpret the data, whilst the other is to predict experimental quantities of novel molecules.

Naïve Bayesian classification is a probabilistic classifier method based on the Bayes' theorem,[113] originally published in the eighteenth century by statistician Thomas Bayes. In its simplest form, a naïve Bayes classifier assumes that the presence or absence of one particular feature is unrelated to the presence or absence of another. That is, that the variables are independent of one another. The Bayes' theorem can be used to model the probability distribution of an output variable, if conditional probabilities of the input variables for a set of classes are known.

The Bayes' theorem can be derived relatively easily through several assumptions. That the conditional probability of an event A, given event B, can be written as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Similarly the probability of event B, given event A, can be written as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

The joint probability, $P(A \cap B)$, is common to both equations, hence after rearranging and equating the remainder we arrive at:

$$P(A|B)P(B) = P(B|A)P(A)$$

This can then be rearranged further to give the most commonly known form of the Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

In chemoinformatics, event B typically represents the presence or absences of a particular molecular feature in a molecule, with the predicted likelihood of event A being the activity of said molecule. Hence, using Bayes' theorem to assess the distribution of features in a dataset with different classes (i.e. active and inactive), allows for the prediction of class membership given certain features in a molecule.

$P(A)$, termed the prior in Bayes' nomenclature, makes an assumption about the likelihood of A in the absence of any additional information. In many cases, information from a given dataset can be used to make an estimate of the prior, so that in situations such as virtual screening, in which compounds are ranked relative to one another, the prior will be identical for every compound, with a uniform prior ($P(A) = 1$) often used for simplicity.

As the probability $P(B)$ may be zero for a particular property not encountered in the training set, this exception needs to be treated differently, as do cases where the sample size is very small. In cases where a feature is only present once in an active molecule, but never in an inactive molecule, the likelihood according to the naïve Bayes' classifier that a new molecule will show that feature is 100%. To overcome such situations the Laplacian or Laplace correction is used, which adds k virtual samples to the dataset with the assumption that in each of them a particular feature will be present.[114]

Bayes' theorem is often employed in classification tasks and is termed naïve Bayesian classification. It gets its name (naïve) from the fact that it only considers the frequency of individual features in the classification, and not their mutual dependence. This assumes that every input variable increases or decreases the confidence in a particular category assignment for a molecule. Classification

analysis can even be performed when the variables are numerical in nature. Whilst nominal attributes are more easily dealt with, there are a number of ways in which these numerical attributes can be processed. One option is the use of binning to assign specific values to a particular bin that spans across the range of values. Bin borders can be at either fixed intervals or set such that bins are equally populated. An alternative option is to assume the molecules follow Gaussian distribution, and to then use particular variable thresholds to classify molecules accordingly.

Whilst other machine learning methods have been shown to occasionally perform better than naïve Bayesian classification,[115] its ability to accommodate knowledge from multiple active compounds has shown it to outperform other commonly used methods which are based on both 2D and 3D features.[116, 117] It's true strength lies in its combination of very good performance, together with its time efficient learning and ability to handle multiple categories of variables efficiently.

The main thrust of naïve Bayesian classification and its use to distinguish between active and inactive compounds is based upon two core assumptions:[118] (i) that the descriptors in the training set are equally important as each other, and (ii) that the descriptors are independent of one another. From this the combined probability of a particular activity classification is obtained by multiplying the individual probabilities. Though, these assumptions are often violated, naïve Bayesian classification is robust to such violations, and tolerant towards noise in the experimental data.[119]

Machine learning approaches such as Bayesian classification have been extensively applied to the field of molecular similarity,[116, 117, 120] as well as other virtual screening and scaffold hopping problems.[18, 50, 78, 121] Even though naïve Bayesian

classification is a relatively recent addition to the arsenal of tools available for chemoinformaticians, it has shown to be of great use when implemented in conjunction with classical modelling techniques, to assist in the rapid virtual screening of large compound libraries in a systematic manner.[122] As mentioned previously the Bayesian classifiers have the added benefit of being able to handle a variety of numerical or binary data, making the addition of new parameters to existing models a relatively straight forward process. As a result, during a drug discovery project these classifiers can evolve with the needs of the projects, going from general models in the lead finding stage, to increasingly precise models during optimisation.

One example involves the use of Bayesian statistics to model both general (multifamily) and specific (single target) kinase inhibitors, to be used in the therapeutic intervention of cancer, inflammatory diseases and diabetes.[54] A generalised model generated using inhibitor data for multiple kinase classes showed meaningful enrichment for several specific kinase targets. The results were used to prioritise compounds for screening, and to optimally select compounds from third party data collections. The observed benefits of this approach was to find compounds that were not structurally related to known actives, and to find novel targets for which there was not enough information available to build a specific kinase model.

Given that classification methods utilise the structures of both the active and inactive compounds, the structural information of the twelve active and seven inactive molecules tested against *Pf*bc$_1$ (table 21.1) were used to generate naïve Bayesian models. It has been found that when only a small number of active data points are available, the performance of a model may suffer as a result,[112] with small being

defined as a dozen or so active compounds.[123] Above this and it has been found that Bayes classifier should give good performance in typical virtual screening situations, as shown in large scale retrospective virtual screening studies.[117] Though there are only twelve compounds active against *Pf*bc$_1$ (and seven inactive), generated models were statistically analysed to ensure definitive model significance. Additionally, as there is coverage of the inactive class this has been found to be beneficial,[124] as otherwise the Bayes classifier may not be able to draw conclusions from a specific feature set, and thus remain undecided about the classification of a new molecule.

The nineteen molecules were each assigned a qualitative label of either 'yes' or 'no', depending on whether they were active of inactive respectively. For each of these nineteen compounds a number of molecular descriptors were calculated using Pipeline Pilot Student Edition v6.1.[40] As we had successfully used a number of physicochemical descriptors for PCA model generation, the same eight molecular descriptors were decided upon for use with naïve Bayesian classification. To refresh these descriptors were: AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings. The Num_Fragments descriptor was also included. It should be noted that additional descriptor sets were considered (i.e. the twenty molecular properties available within KNIME[125]). However, Bayesian models generated using these descriptors were statistically insignificant, thus additional discussion with regard to such descriptor sets will not occur.

With the nineteen compounds in the dataset correctly labelled and their descriptors calculated, it was now time to build a protocol with which to perform Bayesian classification. This was done using KNIME,[125] a dataflow/workflow program similar to Pipeline Pilot. It contains specific components making it perfectly suited

to data mining problems. An initial set of filters were applied to remove any redundant or uninformative descriptors. A low variance filter with the variance upper bound set to zero resulted in all descriptors being removed that contained only a constant value. Num_Fragments was one such descriptor, showing no variance across the molecules. Next a linear correlation component was applied to calculate the measure of correlation between each of the eight remaining descriptors with respect to one another. As all of the variables were numerical in nature, the Pearson's product moment coefficient was used to assess the correlations.[126] The Pearson's product moment coefficient offers a measure of linear correlation between two variables, with values ranging between +1 to indicate a strongly positive correlation, and -1 to indicate a strongly negative one. The correlations between the descriptors can be observed using the correlation matrix shown in table 2.6. Analysis showed which descriptors were redundant, contributing little or no unique information to the overall model. Most of the descriptors appeared to have very little correlation to each other; however, there was a strong correlation between Num_Atoms and Molecular_Weight. Num_Atoms was also fairly highly correlated to Num_H_Acceptors. Such strong correlations were resolved using a correlation filter, with descriptors removed which had a correlation threshold of greater than 0.9. This cut-off did indeed remove the Num_Atoms descriptor, resulting in only seven physicochemical descriptors for use in Bayesian model development.

Table. 2.6 Correlation matrix between the physicochemical descriptors in the Bayesian model.

	ALogP	**Molecular_Weight**	**Num_H_Acceptors**	**Num_H_Donors**	**Num_Atoms**	**Num_RotatableBonds**	**Num_Rings**	**Num_AromaticRings**
ALogP		0.52	-0.02	-0.57	0.50	0.38	0.12	0.53
Molecular_Weight			0.72	-0.34	**0.98**	0.61	-0.03	0.28
Num_H_Acceptors				-0.19	0.79	0.60	-0.22	-0.11
Num_H_Donors					-0.35	-0.25	-0.15	-0.52
Num_Atoms						0.67	-0.04	0.29
Num_RotatableBonds							-0.68	-0.12
Num_Rings								0.66
Num_AromaticRings								

Prior to model generation the molecular descriptors had to be normalised. Min-Max normalisation was performed to create a linear transformation of the values across all the molecules for each of the descriptors. This linear transformation involved scaling the range of values for each descriptor, with a maximum and minimum value for each descriptor across the molecules of 1 and 0 respectively.

The naïve Bayesian learner component in KNIME[125] was used to create a Bayesian model to classify compound activity, based on the Gaussian distribution of the normalised descriptors for the nineteen compounds in the training set. The model could then be tested using the naïve Bayes predictor component in KNIME,[125] to predict the class membership of external compounds. To begin with the naïve Bayesian model was applied to the same nineteen compounds used in the training. This was to give a clear idea as to how well the model was performing internally, and to illustrate whether or not it was valid. The results of classification prediction when the actual class membership is known can be represented in a confusion of contingency matrix.[127] In such a matrix, the rows represent the actual classes, with

the columns representing the predicted classes which have been assigned based on the model. The matrix shows the number of classifications which have been made correctly, and also where discrepancies in predictions occur. Table 2.7 gives the confusion matrix for the generated naïve Bayesian model when it was applied to the initial nineteen active and inactive molecules used in model development.

Table. 2.7 Confusion matrix of the naïve Bayesian model. Letters correspond to which variables are used when calculating the Cooper statistics (table 2.8).

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**9** a	**3** b	**12** $a+b$
No	**2** c	**5** d	**7** $c+d$
Marginal totals	**11** $a+c$	**8** $b+d$	**19** $a+b+c+d$

From the confusion matrix it is possible to calculate a number of statistics which are useful for model validation, termed the Cooper statistics.[128] These were originally described in the seventies for the validation of carcinogen screening tests,[129] but have been extensively applied to assess the results of classification and Bayesian approaches.[128, 130] The Cooper statistics associated with the confusion matrix in table 2.7 are shown in table 2.8, as is the description and formula required for their calculation.

Table. 2.8 Cooper statistics for the naïve Bayesian model. Letters correspond to which variables from the confusion matrix (table 2.7) are used when calculating the Cooper statistics.

Statistic	Formula	Value	Definition
Sensitivity (True positive rate)	$\frac{a}{(a+b)}$	0.75	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	$\frac{d}{(c+d)}$	0.71	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	$\frac{(\boldsymbol{a+d})}{(\boldsymbol{a+b+c+d})}$	**0.74**	**Fraction of chemicals correctly assigned**
Positive prediction	$\frac{a}{(a+c)}$	0.82	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	$\frac{d}{(b+d)}$	0.63	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	$\frac{c}{(c+d)}$ $1-specificity$	0.29	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	$\frac{b}{(a+b)}$ $1-sensitivity$	0.25	Fraction of active chemicals that are falsely assigned to be non-active

Perhaps the most important and relevant statistic from table 2.8 is the accuracy or concordance. For the naïve Bayesian model this was shown to be 0.74, that is that 74% of the time, the Bayesian model correctly predicted the activity class for the compounds used to build the model, with fourteen of the nineteen compounds classified correctly. It has previously been put forward that for standalone classification models, the Cooper statistics should be significantly greater than 50%.[128] This is clearly obeyed for this model, and there are also a number of other encouraging statistics, i.e. that 82% of the active compounds were correctly assigned as active. What is also encouraging is that the percentage of false negatives (25%) is reasonably small given the size of the dataset.

Whilst these statistics are all encouraging with regard to the naïve Bayesian model, it is still necessary to perform external validation. External validation refers to the application of the model to chemical structures which were not used in the training set to generate the model, yet whose actual activity classes are known (test set).

External validation has proven to be one of the most stringent forms of validation, yet the test set compounds must all lie within a sufficient domain of applicability in order to be useful. To test the model for its external validation, the dataset was partitioned using stratified sampling, such that 70% of the molecules (13 compounds) were in the training set, and 30% (6 compounds) in the test set, a splitting pattern which has been supported by the literature.[128] A new model was built using the training set, and then applied to the training set molecules to give the confusion matrix and Cooper statistics observed in tables 2.9 and 2.10 respectively. As can be seen, the model performs strong internally, with an accuracy of 77%.

Table. 2.9 Confusion matrix of the naïve Bayesian model built using 70% of the molecules, and applied to the training set.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**6**	**2**	**8**
No	**1**	**4**	**5**
Marginal totals	**7**	**6**	**13**

Table. 2.10 Cooper statistics for the naïve Bayesian model built using 70% of the molecules, and applied to the training set.

Statistic	**Value**	**Definition**
Sensitivity (True positive rate)	0.75	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.80	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.77**	**Fraction of chemicals correctly assigned**
Positive prediction	0.86	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.67	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.20	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.25	Fraction of active chemicals that are falsely assigned to be non-active

The true test of its predictive ability lies in its potential to correctly predict the activity class of compounds not used during model development (i.e. those in the training set). Thus the model was applied to the test set of compounds (the other 30% of the dataset) and the same analysis performed. The confusion matrix and Cooper statistics for the external set of compounds is shown in tables 2.11 and 2.12 respectively.

Table. 2.11 Confusion matrix of the naïve Bayesian model built using 70% of the molecules, and applied to the test set.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**3**	**1**	**4**
No	**1**	**1**	**2**
Marginal totals	**4**	**2**	**6**

Table. 2.12 Cooper statistics for the naïve Bayesian model built using 70% of the molecules and applied to the test set.

Statistic	**Value**	**Definition**
Sensitivity (True positive rate)	0.75	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.50	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.67**	**Fraction of chemicals correctly assigned**
Positive prediction	0.75	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.50	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.50	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.25	Fraction of active chemicals that are falsely assigned to be non-active

Whilst the externally accuracy of the model fell to 67%, this is still significantly higher than the aforementioned 50% significance level, and as such can be considered as significant and potentially useful. From this it was possible to deduce that a sound model had indeed been developed, and whilst the statistics could be

considered a little low, this may be attributed to the fact that the test set consisted of only six molecules. Additional confidence and support for the model was attained by using the leave-one-out (LOO) cross validation procedure.[128] This involved employing n training sets in which one molecule had been excluded from the original dataset. A total of n models were then developed by using each of the training sets containing $n-1$ objects, with the developed model used to predict the activity class of the excluded compound. For classification models such as this, the cross validated Cooper statistics can be calculated to test the overall accuracy of the models developed. Tables 2.13 and 2.14 show the confusion matrix and Cooper statistics respectively for the LOO cross validation of the naïve Bayesian model. Cross validation was performed one hundred times and an average across all the results taken.

Table. 2.13 Confusion matrix of one hundred iterations of the LOO cross validation calculation for the naïve Bayesian model.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**11**	**1**	**12**
No	**6**	**1**	**7**
Marginal totals	**17**	**2**	**19**

Table. 2.14 Cooper statistics of one hundred iterations of the LOO cross validation calculation for the naïve Bayesian model.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.92	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.14	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.63**	**Fraction of chemicals correctly assigned**
Positive prediction	0.65	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.50	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.86	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.083	Fraction of active chemicals that are falsely assigned to be non-active

The overall accuracy across one hundred iterations of the LOO cross validation was 63%, which was in good agreement with the Cooper statistics of external validation. All things considered it appeared that a fairly significant and robust model has indeed been developed. However, before the model was applied to the ZINC lead like library of compounds, an additional approach was considered to try and better represent the initial data, and thus develop a more significant and accurate model.

The synthetic minority oversampling technique (SMOTE) can be used to oversample the input data, by adding artificial rows to enrich the data.[131] When using learning algorithms it can sometimes be useful to have an equal class distribution of molecules in order to achieve a good classification performance. In cases where there is an unbalance in the input data, for instance if there are only a few objects in the active class but many in the inactive, this would leave the active class of compounds underrepresented, and thus the dataset skewed in favour of the inactive compounds. SMOTE can therefore be used to counteract such discrepancies, and adjust the class distribution by adding artificial rows for the active compounds. The SMOTE algorithm works by creating a synthetic object which is an extrapolation between a real object, and one of its nearest neighbours in a particular class. It picks

a point along the line between these two objects and determines the attributes of the new object, based on this randomly chosen point. This oversampling of the data is used to enrich the data and hopefully develop more meaningful and useful models.

The SMOTE component in KNIME[125] was used to correct the underrepresentation of the inactive class in the initial dataset (twelve active; seven inactive). The SMOTE algorithm picked an object from the inactive class, and then randomly selected one of its nearest neighbours, drawing the new synthetic object along the line between the sample and the nearest neighbour. The data was oversampled by the minority class, so synthetic objects were only added to the inactive class. Five synthetic objects were created, resulting in twelve compounds in each activity class.

A new naïve Bayesian model was developed using the twenty four molecules (twelve active; twelve inactive) in the training set. This model was built as described previously, and then tested on the entire training set to see how well it performed internally. The confusion matrix and Cooper statistics for this model are shown in tables 2.15 and 2.16 respectively.

Table. 2.15 Confusion matrix of naïve Bayesian model built after using SMOTE.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**9**	**3**	**12**
No	**3**	**4**	**7**
Marginal totals	**12**	**7**	**19**

Table. 2.16 Cooper statistics of naïve Bayesian model built after using SMOTE.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.75	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.57	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.68**	**Fraction of chemicals correctly assigned**
Positive prediction	0.75	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.57	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.43	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.25	Fraction of active chemicals that are falsely assigned to be non-active

It can be seen that when using SMOTE, the accuracy of the naïve Bayesian model fell from 74%, to 68%. Though not a huge drop, it was still interesting to note that oversampling the inactive class led to a less significant model. To validate this conclusion further, the dataset was partitioned as before so that 70% of the molecules were in the training set, and the other 30% in a test set. A new model was developed for the training set and then tested internally against the same set of compounds, with the confusion matrix and Cooper statistics shown in tables 2.17 and 2.18 respectively.

Table. 2.17 Confusion matrix for the naïve Bayesian model built after using SMOTE on 70% of the molecules and applied to the training set.

Actual \ Predicted	Yes	No	Marginal totals
Yes	**5**	**3**	**8**
No	**1**	**4**	**5**
Marginal totals	**6**	**7**	**13**

Table. 2.18 Cooper statistics for the naïve Bayesian model built after using SMOTE on 70% of the molecules and applied to the training set.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.63	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.80	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.69**	**Fraction of chemicals correctly assigned**
Positive prediction	0.83	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.57	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.20	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.38	Fraction of active chemicals that are falsely assigned to be non-active

Initially these results looked encouraging, reporting an accuracy of 69%. However, when the model was applied to the test set the Cooper statistics demonstrated the unsuitability of the model, with an accuracy on only 33%, demonstrated by the confusion matrix and Cooper statistics in tables 2.19 and 2.20 respectively. This value is much lower than the minimum 50% accuracy which was discussed earlier.

Table. 2.19 Confusion matrix for the naïve Bayesian model built after using SMOTE on 70% of the molecules and applied to the test set.

Actual \ Predicted	Yes	No	Marginal totals
Yes	**1**	**3**	**4**
No	**1**	**1**	**2**
Marginal totals	**2**	**4**	**6**

Table. 2.20 Cooper statistics for the naïve Bayesian model built after using SMOTE on 70% of the molecules and applied to the test set.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.25	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.50	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.33**	**Fraction of chemicals correctly assigned**
Positive prediction	0.50	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.25	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.50	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.75	Fraction of active chemicals that are falsely assigned to be non-active

Final support for the unsuitability of SMOTE in this example was provided via LOO cross validation of the model. One hundred iterations of the loop were performed, with an average accuracy of only 50%, as represented by the confusion matrix and Cooper statistics in tables 2.21 and 2.22 respectively.

Table. 2.21 Confusion matrix of one hundred iterations of the LOO cross validation calculation for the naïve Bayesian model built after using SMOTE on 70% of the molecules.

Actual \ Predicted	Yes	No	Marginal totals
Yes	**11**	**1**	**12**
No	**11**	**1**	**12**
Marginal totals	**22**	**2**	**24**

Table. 2.22 Cooper statistics of one hundred iterations of the LOO cross validation calculation for the naïve Bayesian model built after using SMOTE on 70% of the molecules.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.92	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.083	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.50**	**Fraction of chemicals correctly assigned**
Positive prediction	0.50	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.50	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.92	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.083	Fraction of active chemicals that are falsely assigned to be non-active

The results supported the decision not to use SMOTE, as the raw data alone produced a more accurate and significant model. The consensus of this investigation reinforced the decision to simply move forward using the original naïve Bayesian model which was developed from only the original compounds, and the seven physicochemical descriptors described (tables 2.7 and 2.8).

The significant Bayesian classification model was applied to the ZINC lead like library of compounds. The physicochemical descriptors for the 2.7 million compounds were calculated in Pipeline Pilot Student Edition v6.1,[40] and then normalised in KNIME[125] according to the normalisation procedure used during model development. The Bayesian model was applied to ZINC, and for each molecule a Bayesian probability calculated. This probability was based upon the relationship observed between the molecular descriptors and the associated activity classes. The resulting probabilities allowed for each molecule to be considered as either active or inactive, depending upon which side of the probability threshold it fell into. It is useful to keep in mind that when using the Bayesian classification technique, the output variable (the probability) is trained in a binary manner, so that

the classification which is made only represents the likelihood of a molecule showing that particular character.[112] Compounds with a probability value less than 0.5 were classified as inactive, whilst those with a probability value greater than 0.5 were classified as active.

In Pipeline Pilot Student Edition v6.1[40] the compounds were filtered to remove any which were defined as inactive. From this it was found that 725,190 compounds from ZINC had been classified as active. The number of hits from this method was considerably higher than from previous methods, suggesting that almost a third of the compounds in the ZINC lead like library were active. This is clearly very unlikely. However, despite there being such a large number of hits, the results would still enrich the selection of compounds via the consensus study. Additionally, the probability of an active classification may also influence the results and will be discussed further in Chapter III. Naïve Bayesian classification was performed according to the *'Naïve Bayesian Classification Protocol'* as described in the Experimental Chapter.

2.3.6 Decision Tree Analysis

Decision tree analysis is similar to Bayesian classification, in that it considers the chemical structures and properties of both the active and inactive compounds. Though some interpretable information can be garnered from Bayesian classification, its uses fall mainly in its predictive ability, rather than what it tells us about a particular dataset in terms of its chemical properties. Decision tree analysis is very much in contrast to this, as it consists of developing a set of rules that provide a means of associating a molecules features or descriptor values, with a particular property of interest, such as biological activity.[14] A decision tree is commonly

depicted as a tree like structure, with each node corresponding to a specific rule. These rules may correspond to the presence or absence of a particular chemical feature, or to a specific threshold value for a descriptor.

An example of decision tree analysis is shown in figure 2.15. This decision tree was developed for a series of fifty active and twelve inactive sumazole and isomazole analogues. The active molecules were shown to have inotropic properties that can increase the force of heart muscle contraction without increasing its rate, and thus have found use in cardiac failure to increase heart muscle efficiency.[132] The decision tree could be used to classify unknown molecules by moving down the decision tree, answering each of the rules until a terminal node was reached. The terminal node was then used to assign the molecule into an appropriate class. This particular dataset is of interest because the active compounds are surrounded in the descriptor space by the inactive molecules, which may ordinarily cause problems when using regression methods, but here were found to be beneficial. The decision tree can be used more simply to determine which features appear to give rise to activity, and which lead to inactivity.

Fig. 2.15 Example of decision tree analysis for a series of sumazole and isomazole analogues for their inotropic properties. (M. A-Razzak and R. C. Glen, *Journal of Computer-Aided Molecular Design*, 1992, 6, 349-383.)

Another example shows the development of models in order to predict the likelihood of diabetes mellitus in patients.[133] Diabetes mellitus, or diabetes, is a major global health problem which affects millions of people worldwide. Classification algorithms have been applied in the past to use patient profiles to predict those at greater risk of having diabetes.[134, 135] Owing to the greatly increased amount of data gathered in medical databases, traditional manual analysis has become inadequate, and methods for efficient computer based analysis are now indispensible. In this research, classification methods were used to predict whether patients were likely to suffer from diabetes, based upon symptoms mined from their medical records. A decision tree was built and tested using data from 768 patients. This data contained information such as the individuals' age, blood pressure and body mass index (BMI), with 268 of the patients in the dataset having been diagnosed with diabetes (i.e. the active class). The subsequent decision tree had an accuracy of 89.3%, and when this data was pruned to remove anomalies in the training set such as noise and outliers in

order to avoid over fitting during development, the accuracy improved further to 89.7%.

There are many methods available for the construction of decision trees, but most of them follow the same basic approach, which is to start with the entire dataset and identify the descriptor which gives the best initial split. This enables the dataset to be divided into two or more subsets, with the same procedure then applied to each of these subsets and so on. This is repeated until all of the molecules have been appropriately divided into their distinct classes, or until no more splits of significance can be found. One group of algorithms, such as ID3 (iterative dichotomiser 3) and C4.5, use the results from the field of information theory or entropy, to provide a measure of the uncertainty associated with a random variable to decide which criteria to choose at each step of the decision tree generation.[136] If there are N molecules divided into C_1, C_2, C_3 ... classes, with each class C_i containing n_i molecules, then the information corresponding to this distribution (also called the entropy of the system) is given by equation 2.4, where $p_i = n_i/N$. This equation is used to decide which descriptor to use in order to construct the next rule, that is the one which gives rise to the largest increase in entropy, also termed the gain.

$$I = \sum p_i \ln p_i$$

Eq. 2.4 Equation to calculate the entropy of a system.

Within this thesis the J48 algorithm was used for the construction of decision trees based on the twelve active and seven inactive compounds. The J48 algorithm is an open source Java implementation of the C4.5 algorithm, available from within the WEKA (Waikato Environment for Knowledge Analysis) data mining tool.[137] C4.5 is itself an improved version of ID3, with ID3 having been developed as an

algorithm for use in decision tree generation.[138] However, a disadvantage of ID3 is that it often over fits the training data, which can give rise to decisions trees that are too specific, and which may not be resistant to noise when tested on novel compounds. Another disadvantage of ID3 is that it cannot deal with missing attributes, requiring that all variables have nominal values. It also does not have the means to manage continuous attributes. C4.5 was therefore developed as an extension of ID3, and has been found to prevent this over fitting of the training data by pruning the decision tree when required, thus making it more resistant to noise.[139, 140] The decision trees generated by C4.5 can therefore be used for classification purposes.

The J48 algorithm employs an automatic procedure capable of selecting relevant features from the training data.[133] It is able to cut the poor or non-meaningful branches through an efficient pruning process, as well as being able to handle both continuous and discrete attributes. In handling continuous attributes, J48 creates a threshold, splitting the list into those whose attribute value is above the threshold, and those that are below or equal to it, thus creating rules through an iterative process.

The same nine physicochemical descriptors (AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings; Num_Fragments) which were used for Bayesian classification were again employed for decision tree analysis, and calculated in Pipeline Pilot Student Edition v6.1.[40] As with Bayesian classification the twelve active compounds were labelled ‘yes’ and the seven inactive ‘no’. Decision tree models were developed using the KNIME[125] workflow program. Low variance and correlation filters were first applied to remove the Num_Fragments (no variance) and

Num_Atoms (strong correlation with Molecular_Weight) descriptors respectively, with the descriptor values then normalised using min-max normalisation. For decision tree development it was necessary to convert the activity labels from string variables to nominal variables, thus allowing for classification models to be built. An initial model was built using all the molecules in the training set. The *'J48 (Weka)'* node was used to develop the model, with the resultant model applied to the same training set of compounds using the *'Decision Tree Predictor'* component. This was to review how well the model performed at predicting the correct activity class for the structures used during model development. Success was assessed through the use of a confusion matrix and Cooper statistics, which for this model are shown by tables 2.23 and 2.24 respectively.

Table. 2.23 Confusion matrix of the Decision Tree model with all molecules in the training set.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**10**	**2**	**12**
No	**0**	**7**	**7**
Marginal totals	**10**	**9**	**19**

Table. 2.24 Cooper statistics of the Decision Tree model with all molecules in the training set.

Statistic	**Value**	**Definition**
Sensitivity (True positive rate)	0.83	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	1.00	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.90**	**Fraction of chemicals correctly assigned**
Positive prediction	1.00	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.78	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.00	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.17	Fraction of active chemicals that are falsely assigned to be non-active

The Cooper statistics were found to be highly favourable, showing the model to have an accuracy of 90%. However, when the use of SMOTE was investigated with the minority class being oversampled (as with the naïve Bayesian classification), the accuracy of the model fell to 79%. SMOTE was therefore not considered further as it was proven to worsen the success of decision tree analysis. To validate this model further before it was applied to the ZINC lead like library, the dataset was partitioned using stratified sampling, such that 70% of the molecules were now in a training set, and the other 30% in a test set. A new decision tree was developed for this training set, and the model applied to the training set. The confusions matrix and Cooper statistics for this model are shown in tables 2.25 and 2.26 respectively.

Table. 2.25 Confusion matrix of the Decision Tree model built using 70% of the molecules and applied to the same training set.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**6**	**2**	**8**
No	**0**	**5**	**5**
Marginal totals	**6**	**7**	**13**

Table. 2.26 Cooper statistics of the Decision Tree model built using 70% of the molecules and applied to the same training set.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.75	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	1.00	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.85**	**Fraction of chemicals correctly assigned**
Positive prediction	1.00	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.71	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.00	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.25	Fraction of active chemicals that are falsely assigned to be non-active

Internally the model looked to perform strongly, with an accuracy of 85%. However, to assess its external validity the model was applied to the test set of compounds, with the results reported in the confusions matrix and Cooper statistics of tables 2.27 and 2.28 respectively.

Table. 2.27 Confusion matrix of the Decision Tree model built using 70% of the molecules and applied to the test set of compounds.

Actual \ Predicted	**Yes**	**No**	**Marginal totals**
Yes	**2**	**2**	**4**
No	**0**	**2**	**2**
Marginal totals	**2**	**4**	**6**

Table. 2.28 Cooper statistics of the Decision Tree model built using 70% of the molecules and applied to the test set of compounds.

Statistic	**Value**	**Definition**
Sensitivity (True positive rate)	0.50	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	1.00	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.67**	**Fraction of chemicals correctly assigned**
Positive prediction	1.00	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.50	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.00	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.50	Fraction of active chemicals that are falsely assigned to be non-active

Though the accuracy dropped to 67% when tested externally, the minimum criteria that the Cooper statistic for accuracy should exceed at least 50% was satisfied.[128] Finally, the model built using all nineteen compounds was assessed using the LOO cross validation procedure, and measured across one hundred iterations. The confusion matrix and Cooper statistics for the LOO cross validation of the model are shown in tables 2.29 and 2.30 respectively.

Table. 2.29 Confusion matrix of one hundred iterations of the LOO cross validation calculation for the Decision Tree model built using all molecules.

Actual \ Predicted	Yes	No	Marginal totals
Yes	**11**	**1**	**12**
No	**6**	**1**	**7**
Marginal totals	**17**	**2**	**19**

Table. 2.30 Cooper statistics of one hundred iterations of the LOO cross validation calculation for the Decision Tree model built using all molecules.

Statistic	Value	Definition
Sensitivity (True positive rate)	0.92	Fraction of active chemicals correctly assigned
Specificity (True negative rate)	0.14	Fraction of non-active chemicals correctly assigned
Concordance or accuracy	**0.63**	**Fraction of chemicals correctly assigned**
Positive prediction	0.65	Fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative prediction	0.50	Fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False positive (over-classification) rate	0.86	Fraction of non-active chemicals that are falsely assigned to be active
False negative (under-classification) rate	0.083	Fraction of active chemicals that are falsely assigned to be non-active

The accuracy was found to be 63% and sufficiently meet the required criteria. Thus it was reasonable to conclude that a validated model had been developed and was ready to be applied to the ZINC lead like library. However, prior to its application, the models decision tree as shown by figure 2.16 was inspected, as one of the useful features of decision tree analysis is how readily interpretable the results are.

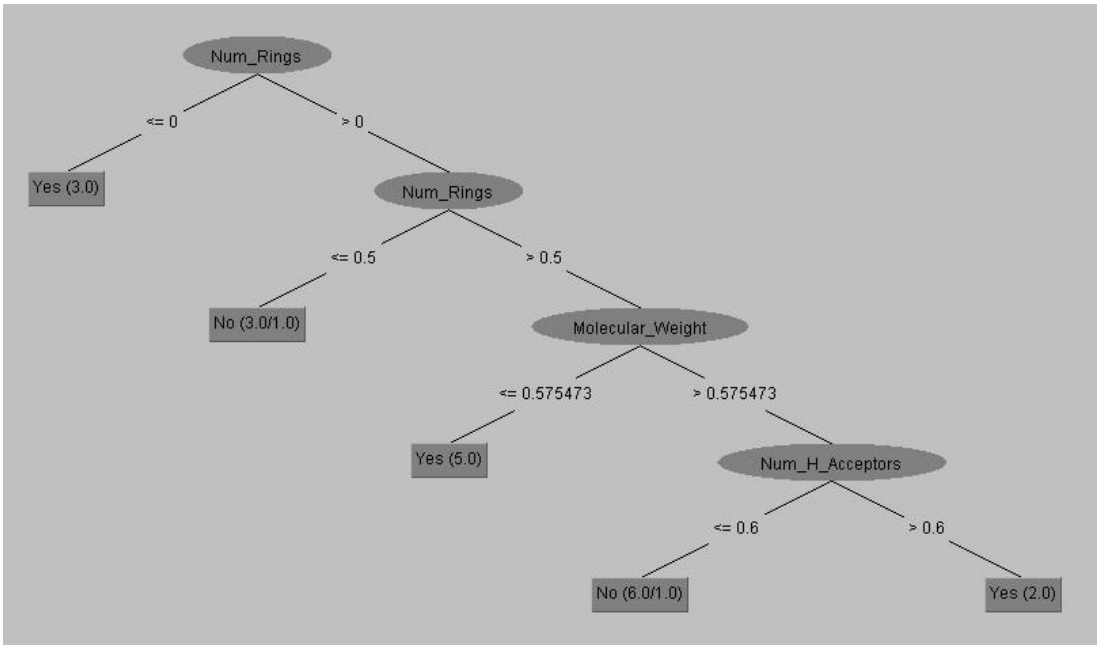


Fig. 2.16 Illustration of the decision tree model built using the structures of the twelve active and seven inactive compounds, which was subsequently applied to the ZINC lead like library of compounds. 'Yes' indicates an active classification and 'no' an inactive one.

Figure 2.16 shows the rules/branches of the decision tree together with their associated threshold values. As can be seen, the decision tree comprises of four rules. The first rule, which according to the J48 algorithm best splits the initial dataset, is associated with the Num_Rings descriptor. It is important to remember that the descriptor values, and thus the threshold values, all refer to the normalised data, for which the spread of the values across each descriptor have a mean of zero and a standard deviation of one. The first rule states that if the normalised Num_Rings descriptor value is less than or equal to zero, then the compound is given a 'yes' (active) classification. Three compounds had a Num_Rings value less than or equal to zero, and were therefore classified as active. Next to the classification in the decision tree there are usually two numbers in brackets. The first is the number of compounds which terminated at this node based on the prior rule, and the second is the number of molecules which were classified incorrectly. Thus for this particular rule, all three compounds were classified correctly as active.

Moving down the decision tree, those compounds which had a Num_Rings descriptor value greater than zero passed to the next branch, which was again concerned with the Num_Rings descriptor. If the normalised descriptor value was less than or equal to 0.5 it was classified as 'no' (inactive). Three compounds were classified as inactive, however, one was misclassified. Given that the overall accuracy of the model was 90%, the odd misclassification was to be expected. Several of the active and inactive compounds have either three or four rings, which perhaps makes this rule slightly limiting, as it does not consider enough variance in the data to fully distinguishing between all active and inactive compounds.

Molecules with a normalised Num_Rings value greater than 0.5 passed to the next rule, with those having a normalised Molecular_Weight descriptor value equal to or less than 0.575473 classified as 'yes' (active). This rule was correct for all of the compounds. The remaining compounds were assessed according to their normalised Num_H_Acceptors descriptor values, so that if it was less than or equal to 0.6, they were classified as inactive, or active if it was greater than 0.6. The former parameter was correct for five of the six molecules, with the later correct for both molecules which terminated here.

With the model now validated and well understood, it was applied to the 2.7 million compounds in the ZINC lead like library. Unlike Bayesian classification, the decision tree analysis did not give a probability for a particular classification, but simply either a 'yes' or 'no' prediction as to whether the compounds may be active or inactive. All the compounds in ZINC were assigned to a particular class, based on their molecular descriptor values according to the decision tree shown in figure 2.16. The results were filtered in Pipeline Pilot Student Edition v6.1[40] to remove any

compounds assigned as inactive, resulting in 1,838,020 compounds classified as active. This number of hits is considerable larger than even that of Bayesian classification, with almost 68% of the chemical library classified as potentially active. Again this would be wildly optimistic if the results were being considered in isolation, but would still act to support the consensus study. Decision tree analysis was performed according to the *'Decision Tree Analysis Protocol'* as described in the Experimental Chapter.

2.4 Merging of the Results

With hits identified from the ZINC lead like library[31, 32] using the six different LBVS methods that were performed in parallel, the results had to be merged prior to consensus analysis. Table 2.31 illustrates the number of hits for each of the methods. These were merged using Pipeline Pilot Student Edition v6.1,[40] giving a total of 1,910,378 unique ZINC entries. This meant that around 70% of the molecules in ZINC had been selected by one method or another, and as such were considered to be worthy of further investigations.

Table. 2.31 Number of hits from ZINC for each of the LBVS methods which have been performed in parallel.

LBVS Method	Number of Hits from ZINC
Fingerprint Similarity Searching	11,655
Turbo Similarity Searching	13,771
Bioisostere Substructure Searching	20,319
Principal Component Analysis	5,000
Naïve Bayesian Classification	725,190
Decision Tree Analysis	1,838,020
Merge Results:	**1,910,378**

The next chapter will now discuss the consensus analysis, together with the scoring of the compounds. A host of filtering methods were employed to identify the most

promising lead like candidates, as well as diversity analysis to ultimately select a number of compounds for purchase and later biological testing.

2.5 References

1. A. R. Crofts, *Annu. Rev. Physiol.*, 2004, **66**, 689-733.
2. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
3. M. Fry and M. Pudney, *Biochem. Pharmacol.*, 1992, **43**, 1545-1553.
4. M. W. Mather, E. Darrouzet, M. Valkova-Valchanova, J. W. Cooley, M. T. McIntosh, F. Daldal and A. B. Vaidya, *J. Biol. Chem.*, 2005, **280**, 27458-27465.
5. J. Krungkrai, S. R. Krungkrai, N. Suraveratum and P. Prapunwattana, *Biochem. Mol. Biol. Int.*, 1997, **42**, 1007-1014.
6. A. Farnert, J. Lindberg, P. Gil, G. Swedberg, Y. Berqvist, M. M. Thapar, N. Lindegardh, S. Berezcky and A. Bjorkman, *Br. Med. J.*, 2003, **326**, 628-629.
7. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
8. P. Mitchell, *Journal of Theoretical Biology*, 1976, **62**, 327-367.
9. M. J. Smilkstein, I. Forquer, A. Kanazawa, J. X. Kelly, R. W. Winter, D. J. Hinrichs, D. A. Kramer and M. K. Riscoe, *Mol. Biochem. Parasitol.*, 2008, **159**, 64-68.
10. W. L. Jorgensen, *Science*, 2004, **303**, 1813-1818.
11. W. P. Walters, M. T. Stahl and M. A. Murcko, *Drug Discovery Today*, 1998, **3**, 160-178.
12. J. A. Bikker and L. S. Narasimhan, Editon edn., 2010, vol. 5, pp. 85-124.
13. P. G. Bray, R. E. Martin, L. Tilley, S. A. Ward, K. Kirk and D. A. Fidock, *Mol. Microbiol.*, 2005, **56**, 323-333.
14. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
15. G. M. Maggiora and M. A. Johnson, *INTRODUCTION TO SIMILARITY IN CHEMISTRY*, John Wiley & Sons Inc, New York, 1990.
16. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, *J. Med. Chem.*, 1996, **39**, 3049-3059.
17. Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350-4358.
18. A. Bender and R. C. Glen, *Org. Biomol. Chem.*, 2004, **2**, 3204-3218.
19. S. R. Vasudevan and G. C. Churchill, *Expert. Opin. Drug Discov.*, 2009, **4**, 901-906.
20. O. Dror, D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson, *Journal of Chemical Information and Modeling*, 2009, **49**, 2333-2343.
21. P. M. O'Neill, The University of Liverpool, 2009.
22. A. Schuffenhauer and T. P. Begley, in *Wiley Encyclopedia of Chemical Biology*, John Wiley & Sons, Inc., Editon edn., 2007.
23. B. K. Shoichet, *Nature*, 2004, **432**, 862-865.
24. C. Lipinski and A. Hopkins, *Nature*, 2004, **432**, 855-861.
25. P. Kirkpatrick and C. Ellis, *Nature*, 2004, **432**, 823-823.
26. *Chemical Abstracts Service (CAS)* - http://www.cas.org/cgi-bin/cas/regreport.pl.
27. R. van Deursen and J.-L. Reymond, *ChemMedChem*, 2007, **2**, 636-640.
28. T. Fink and J.-L. Reymond, *Journal of Chemical Information and Modeling*, 2007, **47**, 342-353.
29. J. Rosen, J. Gottfries, S. Muresan, A. Backlund and T. I. Oprea, *Journal of Medicinal Chemistry*, 2009, **52**, 1953-1962.
30. *National Cancer Institute (NCI)* - www.nci.nih.gov.
31. J. J. Irwin and B. K. Shoichet, *Journal of Chemical Information and Modeling*, 2005, **45**, 177-182.
32. S. J. Teague, A. M. Davis, P. D. Leeson and T. Oprea, *Angew. Chem.-Int. Edit.*, 1999, **38**, 3743-3748.
33. T. I. Oprea, A. M. Davis, S. J. Teague and P. D. Leeson, *Journal of Chemical Information and Computer Sciences*, 2001, **41**, 1308-1315.
34. C. A. Lipinski, *Drug Discovery Today: Technologies*, 2004, **1**, 337-341.
35. D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *Journal of Medicinal Chemistry*, 2002, **45**, 2615-2623.
36. P. R. Andrews, D. J. Craik and J. L. Martin, *Journal of Medicinal Chemistry*, 1984, **27**, 1648-1657.
37. J. C. Baber, W. A. Shirley, Y. Gao and M. Feher, *Journal of Chemical Information and Modeling*, 2005, **46**, 277-288.
38. P. Willett, *Drug Discovery Today*, 2006, **11**, 1046-1053.

39. P. Willett, V. Winterman and D. Bawden, *J. Chem. Inf. Comput. Sci.*, 1986, **26**, 36-41.
40. SciTegic, *Pipeline Pilot Student Edition v6.1*, Accelrys, Inc, San Diego, CA, 2007.
41. ChemBioDraw, *ChembridgeSoft*, 1986-2010.
42. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
43. D. Weininger, A. Weininger and J. L. Weininger, *Journal of Chemical Information and Computer Sciences*, 1989, **29**, 97-101.
44. R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Wiley-VCH, 2000.
45. B. D. Christie, B. A. Leland and J. G. Nourse, *Journal of Chemical Information and Computer Sciences*, 1993, **33**, 545-547.
46. G. M. Maggiora and M. A. Johnson, *Concepts and Applications of Molecular Similarity*, Wiley, New York, NY, 1990.
47. M. J. McGregor and P. V. Pallai, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 443-448.
48. L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, *Classification and Regression Trees*, 1984.
49. D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742-754.
50. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *Org. Biomol. Chem.*, 2004, **2**, 3256-3266.
51. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *Journal of Chemical Information and Computer Sciences*, 2002, **42**, 1273-1280.
52. SciTegic, *Pipeline Pilot*, Accelrys, Inc, San Diego, CA, 2000.
53. M. Hassan, R. D. Brown, S. Varma-O'Brien and D. Rogers, *Molecular Diversity*, 2006, **10**, 283-299.
54. X. Y. Xia, E. G. Maliski, P. Gallant and D. Rogers, *Journal of Medicinal Chemistry*, 2004, **47**, 4463-4470.
55. H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107-&.
56. G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner and A. Zell, *Journal of cheminformatics*, 2011, **3**, 3.
57. J. Greene, S. Kahn, H. Savoj, P. Sprague and S. Teig, *Journal of Chemical Information and Computer Sciences*, 1994, **34**, 1297-1308.
58. *MACCS-II Database System, version 1*, Molecular Design Limited, San Leandro, CA, 1984.
59. P. Gedeck, B. Rohde and C. Bartels, *Journal of Chemical Information and Modeling*, 2006, **46**, 1924-1936.
60. M. D. Krasowski, M. G. Siam, M. Iyer, A. F. Pizon, S. Giannoutsos and S. Ekins, *Clin. Chem.*, 2009, **55**, 1203-1213.
61. L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engqvistgoldstein, R. Bukar, K. E. Bauer, H. Dilley and D. M. Rocke, *Chem. Biol.*, 1995, **2**, 107-118.
62. J. N. Weinstein, T. G. Myers, P. M. Oconnor, S. H. Friend, A. J. Fornace, K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. vanOsdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes and K. D. Paull, *SCIENCE*, 1997, **275**, 343-349.
63. H. Briem and I. D. Kuntz, *Journal of Medicinal Chemistry*, 1996, **39**, 3401-3408.
64. H. Briem and U. F. Lessel, *Perspect. Drug Discov. Design*, 2000, **20**, 231-244.
65. M. Rarey and J. S. Dixon, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 471-490.
66. P. Willett, *Biochem. Soc. Trans.*, 2003, **31**, 603-606.
67. V. Kasam, J. Salzemann, M. Botha, A. Dacosta, G. Degliesposti, R. Isea, D. Kim, A. Maass, C. Kenyon, G. Rastelli, M. Hofmann-Apitius and V. Breton, *Malaria Journal*, 2009, **8**, 88.
68. P. W. Manley, N. Stiefl, S. W. Cowan-Jacob, S. Kaufman, J. Mestan, M. Wartmann, M. Wiesmann, R. Woodman and N. Gallagher, *Bioorganic & Medicinal Chemistry*, 2010, **18**, 6977-6986.
69. E. Gundersen, K. Fan, K. Haas, D. Huryn, J. Steven Jacobsen, A. Kreft, R. Martone, S. Mayer, J. Sonnenberg-Reines, S.-C. Sun and H. Zhou, *Bioorganic & Medicinal Chemistry Letters*, 2005, **15**, 1891-1894.
70. R. S. Ferreira, A. Simeonov, A. Jadhav, O. Eidam, B. T. Mott, M. J. Keiser, J. H. McKerrow, D. J. Maloney, J. J. Irwin and B. K. Shoichet, *Journal of Medicinal Chemistry*, 2010, **53**, 4891-4905.
71. A. R. Burns, I. M. Wallace, J. Wildenhain, M. Tyers, G. Giaever, G. D. Bader, C. Nislow, S. R. Cutler and P. J. Roy, *Nat Chem Biol*, 2010, **6**, 549-557.
72. R. Sharma, A. S. Lawrenson, N. E. Fisher, A. J. Warman, A. E. Shone, A. Hill, A. Mbekeani, C. Pidathala, R. K. Amewu, S. Leung, P. Gibbons, D. W. Hong, P. Stocks, G. L.

Nixon, J. Chadwick, J. Shearer, I. Gowers, D. Cronk, S. P. Parel, P. M. O'Neill, S. A. Ward, G. A. Biagini and N. G. Berry, *Journal of Medicinal Chemistry*, 2012, **55**, 3144-3154.
73. A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies, *Journal of Chemical Information and Modeling*, 2009, **49**, 108-119.
74. R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903-911.
75. P. Willett, *QSAR Comb. Sci.*, 2006, **25**, 1143-1152.
76. J. Hert, P. Willett and D. J. Wilton, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 1177-1185.
77. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *Journal of Medicinal Chemistry*, 2005, **48**, 7049-7054.
78. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *Journal of Chemical Information and Modeling*, 2006, **46**, 462-470.
79. E. J. Gardiner, V. J. Gillet, M. Haranczyk, J. Hert, J. D. Holliday, N. Malim, Y. Patel and P. Willett, *Statistical Analysis and Data Mining*, 2009, **2**, 103-114.
80. J. W. Davies, M. Glick and J. L. Jenkins, *Curr. Opin. Chem. Biol.*, 2006, **10**, 343-351.
81. I. Langmuir, *Journal of the American Chemical Society*, 1919, **41**, 1543-1559.
82. G. L. Patrick, *An Introduction to Medicinal Chemistry*, Oxford University Press, 2005.
83. A. M. Wassermann and J. Bajorath, *Future Medicinal Chemistry*, 2011, **3**, 425-436.
84. K. Kim, J. Kang, S. Kim, S. Choi, S. Lim, C. Im and C. Yim, *Archives of Pharmacal Research*, 2007, **30**, 570-580.
85. D. B. Longley, D. P. Harkin and P. G. Johnston, *Nat Rev Cancer*, 2003, **3**, 330-338.
86. A. Bondi, *The Journal of Physical Chemistry*, 1964, **68**, 441-451.
87. I. Carvalho, Á. D. L. Borges and L. S. C. Bernardes, *Journal of Chemical Education*, 2005, **82**, 588.
88. S. R. Langdon, P. Ertl and N. Brown, *Molecular Informatics*, 2010, **29**, 366-385.
89. K. R. Kim, H. R. Moon, A.-Y. Park, M. W. Chun and L. S. Jeong, *Bioorganic & Medicinal Chemistry*, 2007, **15**, 227-234.
90. A. J. Morrison, J. M. Adam, J. A. Baker, R. A. Campbell, J. K. Clark, J. E. Cottney, M. Deehan, A.-M. Easson, R. Fields, S. Francis, F. Jeremiah, N. Keddie, T. Kiyoi, D. R. McArthur, K. Meyer, P. D. Ratcliffe, J. Schulz, G. Wishart and K. Yoshiizumi, *Bioorganic & Medicinal Chemistry Letters*, 2011, **21**, 506-509.
91. J. Adam, P. M. Cowley, T. Kiyoi, A. J. Morrison and C. J. W. Mort, Editon edn., 2006, vol. 44, pp. 207-329.
92. OpenEye, *BROOD version 1.1.2; http://www.eyesopen.com/brood*, Accessed October 2011.
93. X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *Journal of Chemical Information and Computer Sciences*, 1998, **38**, 511-522.
94. H. Jhoti, A. Cleasby, M. Verdonk and G. Williams, *Curr. Opin. Chem. Biol.*, 2007, **11**, 485-493.
95. F. Fogolari, A. Brigo and H. Molinari, *Journal of Molecular Recognition*, 2002, **15**, 377-392.
96. *Brood: Fragment replacement for Medicinal Chemistry*, OpenEye Scientific Softward, Inc.
97. K. Pearson, *Philos. Mag.*, 1901, **2**, 559-572.
98. A. Palmeira, F. Rodrigues, E. Sousa, M. Pinto, M. H. Vasconcelos and M. X. Fernandes, *Chemical Biology & Drug Design*, 2011, **78**, 57-72.
99. C. F. Higgins, *Current Opinion in Cell Biology*, 1993, **5**, 684-687.
100. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Research*, 2008, **36**, D901-D906.
101. J. N. Weinstein, K. W. Kohn, M. R. Grever, V. N. Viswanadhan, L. V. Rubinstein, A. P. Monks, D. A. Scudiero, L. Welch, A. D. Koutsoukos, A. J. Chiausa and a. et, *SCIENCE*, 1992, **258**, 447-451.
102. A. D. Koutsoukos, L. V. Rubinstein, D. Faraggi, R. M. Simon, S. Kalyandrug, J. N. Weinstein, K. W. Kohn and K. D. Paull, *Statistics in Medicine*, 1994, **13**, 719-730.
103. L. M. Shi, Y. Fan, J. K. Lee, M. Waltham, D. T. Andrews, U. Scherf, K. D. Paull and J. N. Weinstein, *Journal of Chemical Information and Computer Sciences*, 1999, **40**, 367-379.
104. R. R. Meglen, *Journal of Chemometrics*, 1991, **5**, 163-179.
105. M. Calas, G. Cordina, J. Bompart, M. BenBari, T. Jei, M. L. Ancelin and H. Vial, *Journal of Medicinal Chemistry*, 1997, **40**, 3557-3566.
106. P. Broto, G. Moreau and C. Vandycke, *European Journal of Medicinal Chemistry*, 1984, **19**, 61-65.
107. W. P. Walters and M. A. Murcko, *Adv. Drug Deliv. Rev.*, 2002, **54**, 255-271.
108. I. Muegge, *Med. Res. Rev.*, 2003, **23**, 302-321.

109. C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235-249.
110. H. Mishra, N. Singh, T. Lahiri and K. Misra, *Bioinformation*, 2009, **3**, 384-388.
111. P. Willett, J. M. Barnard and G. M. Downs, *Journal of Chemical Information and Computer Sciences*, 1998, **38**, 983-996.
112. A. Bender, *Methods in molecular biology (Clifton, N.J.)*, 2011, **672**, 175-196.
113. T. Bayes, *M.D. computing : computers in medical practice*, 1991, **8**, 157-171.
114. P. Domingos and M. Pazzani, *Mach. Learn.*, 1997, **29**, 103-130.
115. M. Glick, J. L. Jenkins, J. H. Nettles, H. Hitchings and J. W. Davies, *Journal of Chemical Information and Modeling*, 2006, **46**, 193-200.
116. A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 170-178.
117. A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 1708-1718.
118. D. Plewczynski, S. A. H. Spieser and U. Koch, *Journal of Chemical Information and Modeling*, 2006, **46**, 1098-1106.
119. P. Labute, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 1999, 444-455.
120. J. W. Godden and J. Bajorath, *QSAR Comb. Sci.*, 2003, **22**, 487-497.
121. G. Schneider, P. Schneider and S. Renner, *QSAR Comb. Sci.*, 2006, **25**, 1162-1171.
122. A. E. Klon, *Comb. Chem. High Throughput Screen*, 2009, **12**, 469-483.
123. B. Chen, R. F. Harrison, G. Papadatos, P. Willett, D. J. Wood, X. Q. Lewell, P. Greenidge and N. Stiefl, *Journal of Computer-Aided Molecular Design*, 2007, **21**, 53-62.
124. D. Rogers, R. D. Brown and M. Hahn, *J. Biomol. Screen*, 2005, **10**, 682-686.
125. www.knime.org, *KNIME v2.3.3*, 2003-2011.
126. J. L. Rodgers and W. A. Nicewander, *The American Statistician*, 1988, **42**, 59-66.
127. I. E. Frank and J. H. Friedman, *Journal of Chemometrics*, 1989, **3**, 463-476.
128. OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, Paris, 2007.
129. J. A. Cooper, R. Saracci and P. Cole, *Br. J. Cancer*, 1979, **39**, 87-89.
130. M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Science Series 28, Oxford University Press, 2003.
131. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *J. Artif. Intell. Res.*, 2002, **16**, 321-357.
132. M. A-Razzak and R. C. Glen, *Journal of Computer-Aided Molecular Design*, 1992, **6**, 349-383.
133. A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah and N. Y. Yahaya, Editon edn., 2011, vol. 188 CCIS, pp. 537-545.
134. S. F. B. Jaafar and D. M. Ali, *Diabetes mellitus forecast using artificial neural network (ANN)*, Ieee, New York, 2005.
135. J. C. Han, J. C. Rodriguez and M. Beheshti, in *Advances in Software Engineering*, eds. T. Kim, W. C. Fang, C. Lee and K. P. Arnett, Springer-Verlag Berlin, Berlin, Editon edn., 2009, vol. 30, pp. 99-109.
136. J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
137. H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, R. Peter and H. W. Ian, *SIGKDD Explor. Newsl.*, 2009, **11**, 10-18.
138. J. R. Quinlan, *Mach. Learn.*, 1986, **1**, 81-106.
139. S. R. Konda, *A Comparative Evaluation Of Symbolic Learning Methods and Neural Learning Methods*.
140. J. R. Quinlan, *J. Artif. Intell. Res.*, 1996, **4**, 77-90.

Chapter III

Ligand Based Virtual Screening Scoring & Selection

3. Ligand Based Virtual Screening Scoring & Selection 165

3.1 Consensus Scoring 165

3.2 Compound Filtering 174

3.3 Diversity Analysis & Compound Selection 181

3.3.1 Clustering Analysis 182

3.3.1.1 Hierarchical Clustering 184

3.3.1.2 Non-hierarchical Clustering 189

3.4 Visual Inspection 190

3.5 Final Selection/Purchasing Compounds 192

3.6 References 195

3. Ligand Based Virtual Screening Scoring & Selection

Table 3.1 illustrates the number of hits from the ZINC lead like library[1, 2] for each of the LBVS methods which were performed in parallel as described in Chapter II, based on the structures of the compounds tested against *Pf*bc$_1$ (table 2.1). Merging of the hits resulted in 1,910,378 unique entries. Consensus scoring and analysis was used in order to select the most promising candidates from these hits.

Table. 3.1 Number of hits from ZINC for each of the LBVS methods which have been performed in parallel.

LBVS Method	Number of Hits from ZINC
Fingerprint Similarity Searching	11,655
Turbo Similarity Searching	13,771
Bioisostere Substructure Searching	20,319
Principal Component Analysis	5,000
Naïve Bayesian Classification	725,190
Decision Tree Analysis	1,838,020
Merge Results:	**1,910,378**

3.1 Consensus Scoring

Consensus scoring involves preferentially selecting compounds for purchase and biological testing that were identified across multiple virtual screening methods, as oppose to those which were just identified by a single technique. It was hoped that through this, compound selection would be as enriched and informed as possible. Consensus scoring, or data fusion as it is alternatively known, is widely used within protein-ligand docking.[3] Various scoring functions have demonstrated their ability to predict experimental binding affinities,[4-7] but with the introduction of consensus scoring, in which the results from several scoring functions are combined,[8-10] SBVS results have been shown to substantially improve. This has contributed to better enrichments in compound selection over any individual scoring function, allowing for precedence in the data to be spotted. It has been shown that whilst individual

scoring functions can on occasion give sufficient results, a consensus of scoring methods greatly improves them.[11]

Despite consensus scoring receiving wide acceptance from the virtual screening community, there have been relatively few studies dealing with the question as to how consensus scoring actually enriches the datasets. One consideration is that if different scoring functions estimate a property independently, and that that property relates smoothly to an experimentally determined quantity, then the mean from several scoring functions should be a better predictor than each individual score.[12] However, because different scoring functions can have different scales, this makes combining them problematic. This is why with SBVS, ranking methods are often employed to improve virtual screening performance compared with simply combining the raw scores.[3] To do this each compound is first assigned a rank based on its positioning in the primary dataset for each of the different scoring functions. These individual ranks are then combined for each molecule across the function types to give a final rank,[8, 13] with the highly ranking compounds across the consensus shown to have improved activity over individual scoring functions used in isolation.

Several factors may contribute to this improvement in success rates.[3, 14] Firstly, performing multiple screening methods is similar to repeated sampling, so that the mean of the methods will be closer to the true value than any of the individual measurements alone. Secondly, as actives are generally more tightly clustered in chemical space than inactives, multiple sampling is more likely to recover more actives than inactives. Finally, the different methods also seem to agree more on the ranking of actives than on the inactives. This concordance arises given that different scoring functions focus on different aspects of ligand binding, and thus lead to

different false positives. Additionally, it was found that consensus results tend to be more consistent across receptor systems, meaning the user is less dependent on picking the correct method for a particular target of interest.

Whilst consensus scoring has been widely used in SBVS,[9, 12, 15, 16] in LBVS it has mostly been utilised to merge similarity scores and molecular descriptors.[17-21] These consensus decisions are generally based on a combination of the rankings of the different methods, but may lead to complications when different and unrelated algorithms and procedures are used in parallel. This is due to the often dissimilar formatting of the results (i.e. qualitative; quantitative),[14] as observed across the six LBVS methods described in Chapter II. Some of the methods provided continuous values (i.e. similarity searching), whilst others simply gave binary classifications (i.e. Bayesian classification). Care must therefore be taken when merging highly disparate results, in order to obtain the overall improvement in performance which consensus scoring can afford.

The LBVS results (table 3.1) first required normalisation before consensus scoring could be performed, with considerations put in place based on the output from each particular method. The consensus score, or virtual screening score (eq. 3.1), was made up of the summation of the initial scores assigned to a compounds from each of the LBVS methods it was identified by. Firstly, the initial scores for the hits from each LBVS method had to be assigned, and were done so as follows.

The 11,655 and 13,771 compounds which were identified by fingerprint and turbo similarity searching respectively were each assigned a Tanimoto coefficient value, corresponding to their closest active neighbour (table 2.1). Compounds identified from fingerprint similarity searching were scored between 0.7 and 1, and those from

turbo similarity searching between 0.8 and 1. Hits more structurally similar to an active compound were therefore scored more highly. It is important to note that during normalisation and scoring, a hit from a particular method was not able to contribute a score greater than 1. This was so that the final summation of the results would be unbiased, with each method contributing to the overall score.

The 20,319 diverse hits from the bioisostere substructure searching method were each given a score of 1. As each of these compounds contained at least one of the 166 bioisosteres that had been calculated based on the quinolone core, it was deemed only fair that each be allowed to contribute equally to the overall consensus score. It would have been inaccurate to place emphasis onto any particular bioisostere, given that they were all novel fragments, yet to be tested against *Pf*bc$_1$. Thus they were all considered equally as likely to contribute as potential hits.

The 5,000 compounds selected from PCA were those which were closest to known actives in PC space, according to their Euclidean distances. The smaller the distance then the closer the compound was to a known active. The distances for the 5,000 compounds were normalised between 0.5 and 1, such that those with the smallest Euclidean distances (closest to known actives) had the highest scores, and those furthest away the lowest. This normalisation and scoring was performed in Pipeline Pilot Student Edition v6.1,[22] with the histogram shown in figure 3.1 illustrating the distribution of scores.

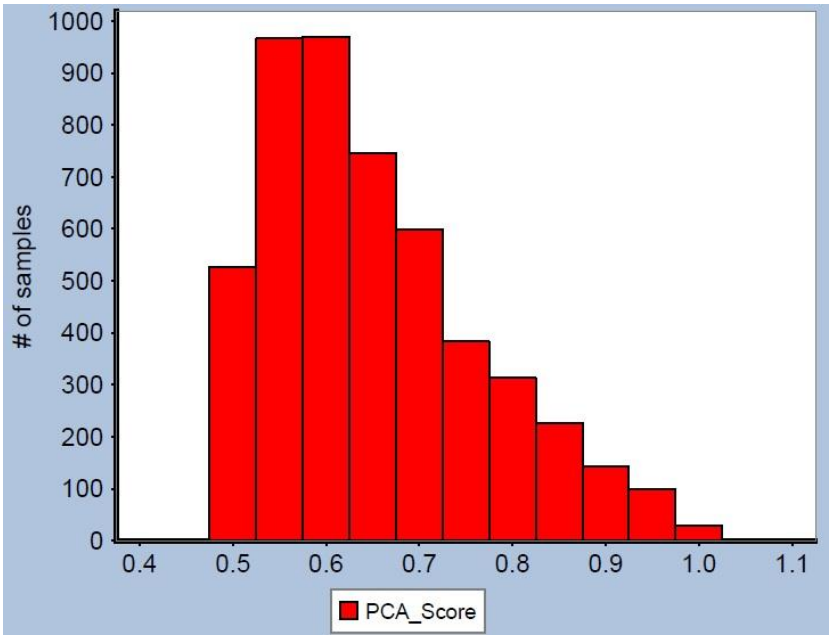


Fig. 3.1 Histogram of the normalised scores for the compounds identified by PCA.

Whilst the number of hits for the previous four methods were all fairly similar to one another, the number of hits from Bayesian classification and decision tree analysis were huge in comparison (table 3.1). However, the results from Bayesian and decision tree could still act to enrich the overall consensus, provided careful limitations were placed upon the results of these methods. The compounds predicted as active from Bayesian were also assigned a probability value, which offered a quantitative means of assessing the classification (i.e. the higher the probability, the greater the support for an active classification). The probability values for the 725,190 compounds classified as active were normalised between 0 and 0.5, with the highest probability scored at 0.5, and the lowest scored at 0. These normalised values represented the new scores for each compound. Though the contributions to the consensus from this method may be smaller compared to other methods (i.e. bioisostere substructure searching), they would still act to strengthen the scores of frequently identified compounds, without biasing the data towards those which occurred less frequently.

Unlike Bayesian classification, decision tree analysis simply classified compounds as either active or inactive, with no probability value to support a particular classification. Each compound was therefore scored equally, which owing to the massive number of compounds classified as active by this method, was set to only 0.25. A score of 0.25 was decided as there were twice as many hits from decision tree analysis compared to Bayesian classification, so it seemed only fair to place half as much emphasis upon the results of this method. However, though the overall contribution of this method compared to the others is relatively small, it still served to strengthen the consensus, offering further support for the more frequently occurring compounds, yet not emphasising any compounds identified by only this method.

With the results normalised and individual scores assigned to each hit, the scores were combined. A term referred to as the virtual screening score (VSS; eq. 3.1) was calculated using Pipeline Pilot Student Edition v6.1,[22] and represents the average value of the scores for a molecule across all methods it was identified by. More specifically, it is the summation of the scores for each molecule, divided by six to account for each of the different LBVS methods. Therefore, compounds which appeared more frequently scored more highly, as appose to those identified by fewer methods.

$$Virtual\ Screening\ Score = \frac{\sum Normalised\ virtual\ screening\ scores}{6}$$

Eq. 3.1 Equation to calculate the virtual screening score (VSS).

Across the 1,910,378 compounds the VSS values varied considerably, having a maximum value of 0.51, and a minimum of $8x10^{-9}$, with the histogram in figure 3.2 representing their distribution. Most compounds scored quite lowly, with only 1,905

having a score greater than or equal to 0.2 (proportion not large enough to be visible on the histogram due to the overwhelming dominance of the lower scoring bins). This observation was not unexpected however, as given the large number of compounds identified by Bayesian classification and decision tree analysis, it was likely that fewer molecules would score highly (i.e. appear in several LBVS methods).

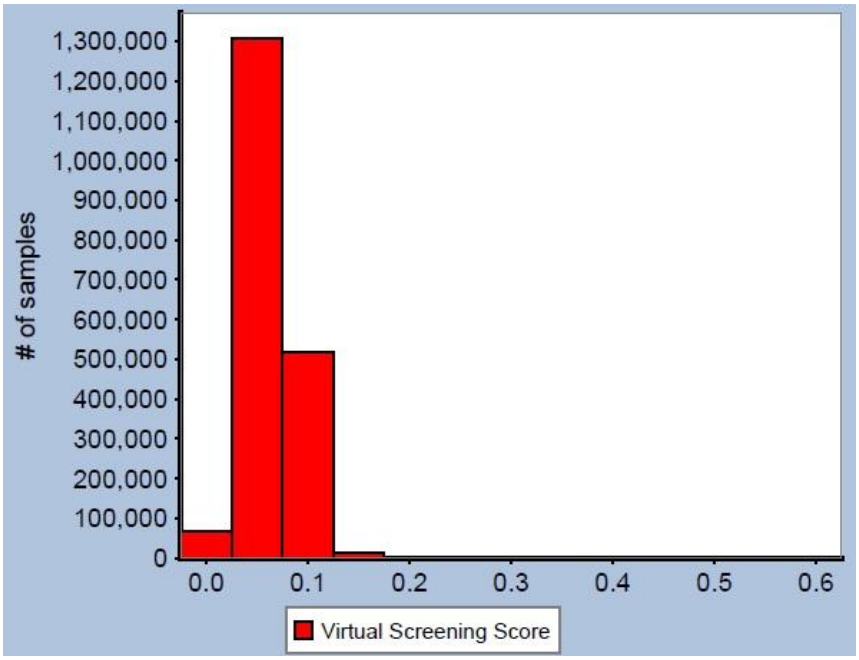


Fig. 3.2 Histogram of the virtual screening scores.

A final score was ultimately calculated for each of the 1,910,378 compounds. This final score was comprised of several properties, with VSS being the main consideration. However, a number of other properties were also considered, including MW, log *P* and log *S*. The importance of MW and log *P* have already been discussed (Chapter I), but log *S* offers a measure of a compounds aqueous solubility, and is often expressed as log units of molar solubility (mol/L). It is an important factor which affects the bioavailability of compounds,[23] with poorly soluble compounds often having bad absorption. These properties were chosen so that compounds were scored, not only according to their frequency from virtual

screening, but also based upon their lead like or drug like potential.[24] It is important to be mindful of a compounds physicochemical properties during drug discovery endeavours, in order to minimise the potential risks of costly, late stage set backs.[25]

$$Final\ Score = 4 * (VSS) + 1 * f(\log S) + 1 * f(\log P) + 2 * f(MW)$$

Eq. 3.2 Equation to calculate the final score for each compound, based upon the VSS and various physicochemical properties.

In accordance with VSS the physicochemical properties had to be normalised so that their contributions ranged between 0 and 1, with more favourable compounds scoring more highly. Throughout the literature there is much debate as to the most appropriate physicochemical descriptor values that best assess lead likeness, so a number of functions were used instead based on a consensus across the literature.[26-28] Additionally, earlier work has shown that the scoring functions illustrated in table 3.2 have been successfully used to identify novel hits active against *Pf*NDH2.[29] These functions were therefore chosen to normalise the physicochemical descriptor values. The functions are also represented graphically in figure 3.3.

Table. 3.2 Scoring function ranges for the molecular properties.

Property	More desirable range	Less desirable range
log S	>-5	<-6
Log P	-1< log P < 4	log P ≤ -2.5; log P ≥ 5.5
MW	<400	>600

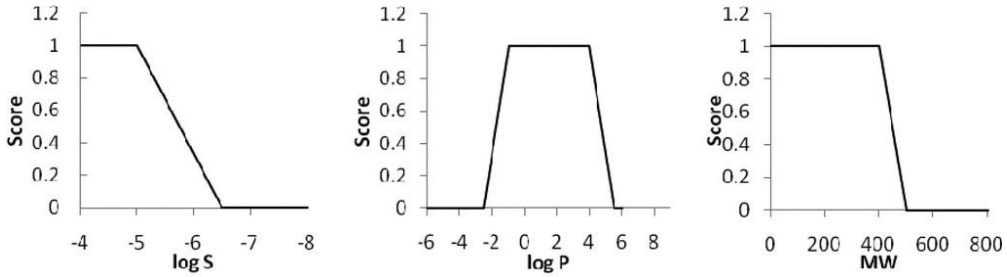


Fig. 3.3 Graphs illustrating the scoring functions for the calculated molecular properties.

The scoring functions described in table 3.2 were implemented using the rules shown in figure 3.4.

log S	$f(\log S) = 1$ if $\log S > -5.0$
	$f(\log S) = 0$ if $\log S < -6.5$
	$f(\log S)$ varies linearly in between
log P	$f(\log P) = 1$ if $\log P > -1.0$ and $\log P < 4.0$
	$f(\log P) = 0$ if $\log P < -2.5$ or $\log P > 5.5$
	$f(\log P)$ varies linearly from 0 to 1 between $-2.5 < \log P < -1.0$
	$f(\log P)$ varies linearly from 1 to 0 between $4.0 < \log P < 5.5$
MW	$f(MW) = 1$ if $MW < 400$
	$f(MW) = 0$ if $MW > 550$
	$f(MW)$ varies linearly between

Fig. 3.4 Functions applied when calculating the final score for the compounds.

When calculating the final score the VSS was multiplied by four. This was to emphasise the importance of the virtual screening results over the other physicochemical properties that make up the rest of the final score. Log *S* and log *P* were both multiplied by one, whereas the function of MW was multiplied by two. Given that in general, the MW and complexity of compounds increases down the drug discovery pipeline, the MW was allowed to contribute more to the overall score compared to log *S* and log *P*. This meant that the final scores were based half on the physicochemical properties of the compounds, whilst the other half was on the results of virtual screening. Figure 3.5 shows a histogram of the final scores for the 1,910,378 compounds. The maximum score was just over 6 and the minimum less than 0.5, though most had a score between 3 and 5. The results are also tabulated across several bins in table 3.3.

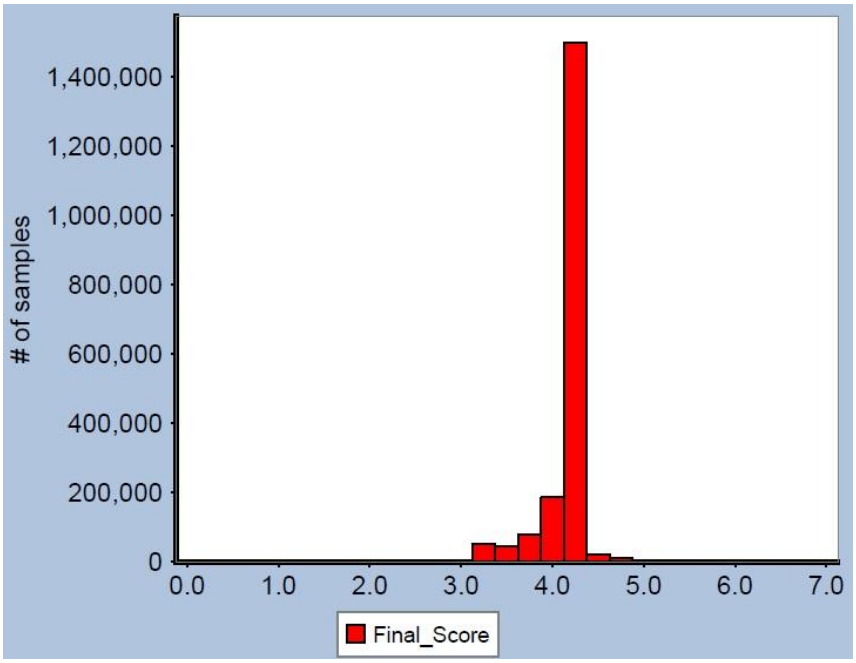


Fig. 3.5 Histogram of the final scores.

Table. 3.3 Frequency of hits in particular Final_Score bins.

Final_Score Ranges	Frequency of Hits
0 to 1	28
1 to 2	501
2 to 3	9,457
3 to 4	239,924
4 to 5	1,660,168
5 to 6	298
6 to 7	2

With each compound from virtual screening scored, a host of filters were applied in order to reduce the number of compounds for consideration, resulting in only those of most interest.

3.2 Compound Filtering

Filtering was performed through a sequential protocol developed in Pipeline Pilot Student Edition v6.1.[22] Lipinski's rules[30, 31] are widely used in drug design as a guideline for compound selection, therefore the dataset was filtered to remove those molecules which violated two or more of its recommendations (≤ 5 H-bond donors; ≤ 10 H-bond acceptors; MW ≤ 500 Da; log P ≤ 5). Though the ZINC lead like library had already undergone some filtering,[2] the Lipinski filter was used as a

precaution to remove potential outliers in the data. As it turned out there were a total of 7,884 compounds which violated two or more of Lipinski's guidelines (1,902,494 compounds remained). Further to this a filter was applied to remove compounds which did not follow Veber's guidelines for oral bioavailability (ten or fewer RBs; PSA less than or equal to 140Å^2).[32] This however did not remove any additional compounds, suggesting that all remaining compounds had acceptable physicochemical profiles according to these widely used methods.

Next chiral molecules were removed. This was to ensure that potential scaffolds for future synthetic optimisation would be as structurally simple and uncomplicated as possible. The enantiomers of chiral compounds can often have different effects as drugs. One example of this is the chiral molecule mefloquine (fig. 3.6), which is a potent inhibitor of the malaria parasite, and a synthetic analogue of quinine (fig. 1.2). However, mefloquine contains two asymmetric carbon atoms, thus having four different stereoisomers. It is currently sold as a racemate of the (*R*,*S*)- and (*S*,*R*)-enantiomers, and is essentially two drugs in one. Plasma concentrations of the (-)-enantiomer are significantly higher than those of the (+)-enantiomer, with the pharmacokinetics between the two enantiomers also significantly different.[33] The (+)-enantiomer has a shorter half life than the (-)-enantiomer, with some research suggesting that it is more effective in treating malaria, as the (-)-enantiomer specifically binds to the adenosine receptors in the central nervous system, which may explain the drugs psychotropic effects.[34] Additionally, achiral molecules are preferred during screening owing to the cost and difficulty of synthesising chiral molecules.[35] In total 850,246 chiral molecules were removed, leaving just over a million (1,052,248) compounds for consideration.

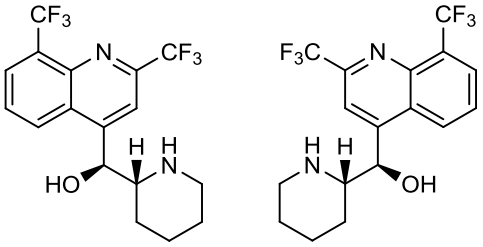


Fig. 3.6 (RS, SR)-(±)-Mefloquine structures.

The next phase of filtering removed the largest proportion of molecules, and was concerned with removing those molecules which contained structural motifs which may potentially have been considered as "non-drug-like", based on precedent across the literature. The first set of motifs were those described by the Rishton fragments.[36, 37] When performing screening it is important to be able to distinguish between promising drug leads and the many false positives which plague screening efforts. False positives are compounds that have activity *in vitro*, which is most likely due to reasons other than hitting the desired target. The Rishton fragments attempt to provide simple chemical guidelines for the evaluation of positives in biochemical screens, with the aim being to select stable, non-covalent ligands, and to eliminate those compounds which may react unfavourably. The early and systematic removal of suspected "nasty" compounds is vital to the success of screening efforts, which is why in this research these filters were applied before compounds were purchased and tested, so as to eliminate any potential difficulties later on. Such fragments may be chemically reactive towards proteins, such as alkylating and acylating agents, and may be prone to hydrolysis, or characteristically reactive towards biological nucleophiles. Figure 3.7 illustrates the twenty five structural motifs encompassed by the Rishton fragments.[36, 37] These reactive functional groups are generally prone to decomposition under hydrolytic conditions, and are reactive towards protein and biological nucleophiles such as glutathione and dithiothreitol. They also exhibit poor stability in serum. It has been stated that compounds which

contain these functional groups should not be submitted for assay in a biochemical screen, without the understanding that they may produce a false positive readout. This generalisation intends to emphasise that reactive compounds that exert their effect in a biochemical screen via covalent bonding are false positives. Regardless of potency, the observed biochemical recording will simply be a consequence of chemical reactivity, not biological activity. Potential antimalarials in this research were expected to inhibit *Pf*bc$_1$ through non-covalent interactions, so it was crucial that chemically reactive compounds be recognised as such as early as possible.

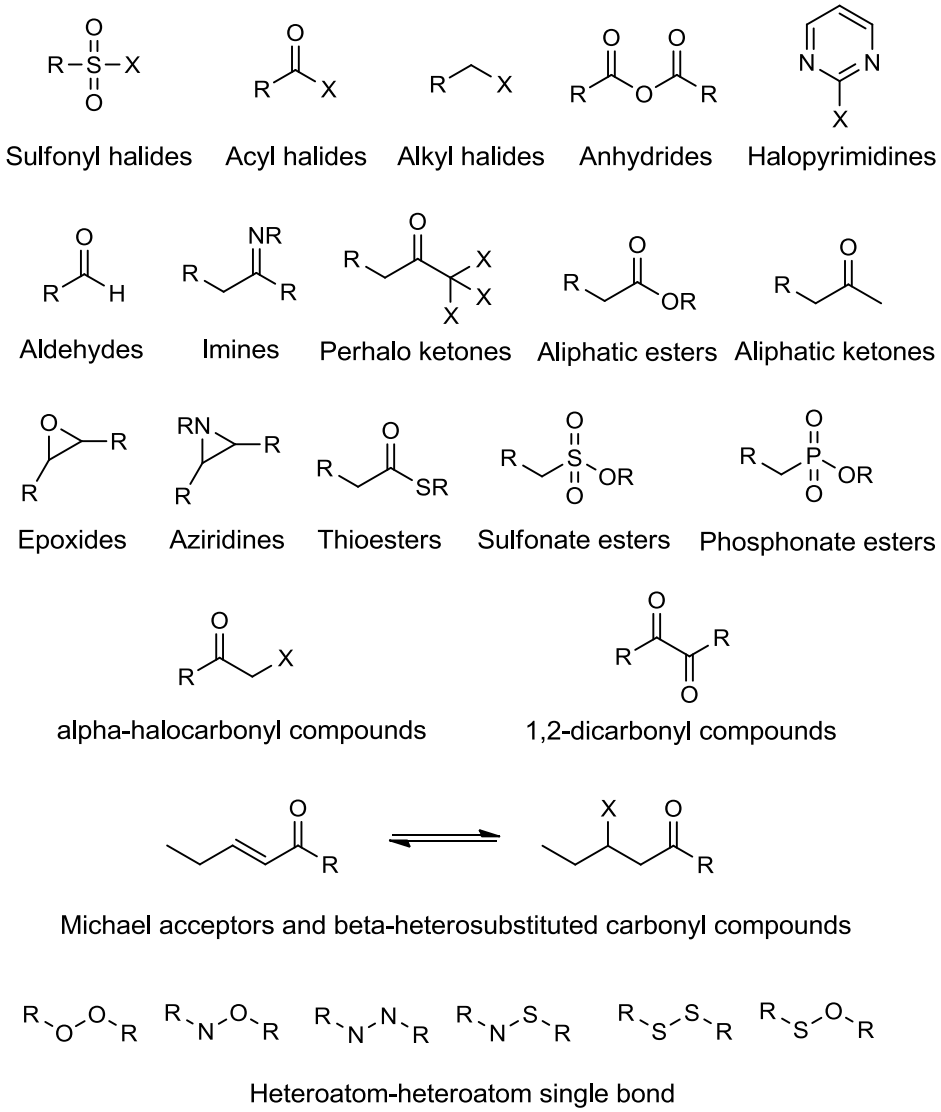
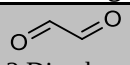
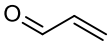
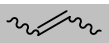
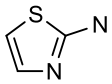
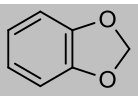
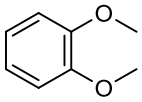
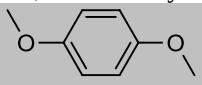
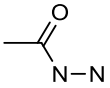
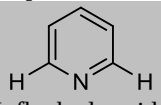
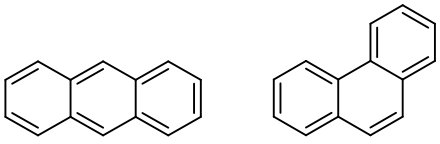


Fig. 3.7 Rishton fragments. (G. M. Rishton, *Drug Discovery Today*, 1997, **2**, 382-384.)

The Rishton fragments were filtered in combination with an additional 274 structural motifs commonly referred to as toxicophores. Toxicophores represent features or groups within a chemical structure that may be responsible for a compounds toxic properties, either directly, or through metabolic activation.[38] A compound often exerts its toxicity through covalent bonding, or via oxidation to cellular macromolecules such as proteins or DNA, thus causing changes to normal cellular biochemistry and physiology, thereby eliciting its toxic effect. These 274 toxicophores were taken from several literature sources, and had previous been used successfully to identify false positives.[39-45] By compiling a list of the fragments presented within these papers it was hoped that the chemical space across which false positives were identified would be expanded. As with the Rishton fragments, these motifs largely remove compounds with specific functional groups that are known to interfere with biochemical assays, or cause problems later in drug development. Though most chemical libraries used for screening contain very few, if any, of the most troublesome functional groups (i.e. aldehydes, epoxides, α-halo ketones), many still contain some problematic motifs. Table 3.4 contains several examples of these toxicophores, with compounds containing these motifs removed during filtering (see appendix for all 274 toxicophores).

Table. 3.4 Examples of chemical fragments used when filtering compounds, together with reasons for their exclusion. (D. J. Huggins, A. R. Venkitaraman and D. R. Spring, *ACS Chem. Biol.*, 2011, **6**, 208-217.)

Chemical Fragment	Reason for Exclusion
 1,2 Dicarbonyl	Metabolically unstable/potential toxicity due to mutagenicity
 α,β-Unsaturated Carbonyl	Prone to reactivity by acting as a Michael acceptor
 Alkene	Metabolically unstable due to epoxidation
 Aminothiazole	Potential toxicity
 Methylenedioxy	Metabolically unstable due to acetal hydrolysis/prone to oxidation yielding reactive quinones
 1,2 Dimethoxy	Prone to oxidation yielding reactive quinones
 1,4 Dimethoxy	Very prone to oxidation yielding reactive quinones
 Acylhydrazide	Metabolically unstable due to acyl hydrolysis
 Acetal	Metabolically unstable due to acetal hydrolysis
 Thiourea	Metabolically unstable due to flavin oxidation/potential non-specific protein binding
 Nitro group	Prone to reduction yielding reactive species/potential hepatocarcinogens
 Aliphatic ketone	Metabolically unstable due to nucleophilic attack
 Unflanked pyridyl	Potential interference with cytochrome P450s due to metal ion coordination
 Anthracene/Phenanthrene	Known DNA intercalation

The remaining 1,052,248 compounds were filtered to remove those which contained any of the Rishton fragments or the literature toxicophores. Again, this filter

removed the largest proportion of molecules (860,056) of all the filtering methods, passing only 192,192 compounds to the next stage.

The last filtering measure was to remove all compounds containing anions and/or cations. Whilst ionic compounds are generally more soluble in water than their neutral counterparts, they tend to be less soluble in organic solvents. Also, compounds must pass through cell membranes in order to reach their targets, and these lipid bilayers are selectively permeable to ions and organic molecules, and control the movements of substances in and out of the cells, generally being permeable to only small, uncharged molecules.[46] As solubility favours charged species, and permeability favours neutral ones, there needs to be a balance between these opposing properties, allowing for a molecule to dissolve, permeate and be absorbed. This filter reduced the dataset further to 127,725 compounds.

The remaining compounds still observed a good spread of final scores, with values ranging between 5.38 and 2.29. A cut-off was placed so that only compounds which exceeded a particular threshold were considered further. This cut-off was decided through consideration of equation 3.2, and the method by which the final score was generated. According to equation 3.2, it would be possible for a compound to have a score of 4 without even taking into account the virtual screening work, provided it was sufficiently small, soluble and had favourable lipophilicity. Though good physicochemical properties were highly desirable, the primary aim was to identify compounds which had been identified across multiple LBVS methods. Therefore the cut-off was placed just above 4 at 4.5. Figure 3.8 shows a histogram of the spread of scores across the 127,725 filtered compounds, together with a line intersecting the point at which the cut-off was placed. In total there were 2,538 compounds which had a final score greater than or equal to 4.5.

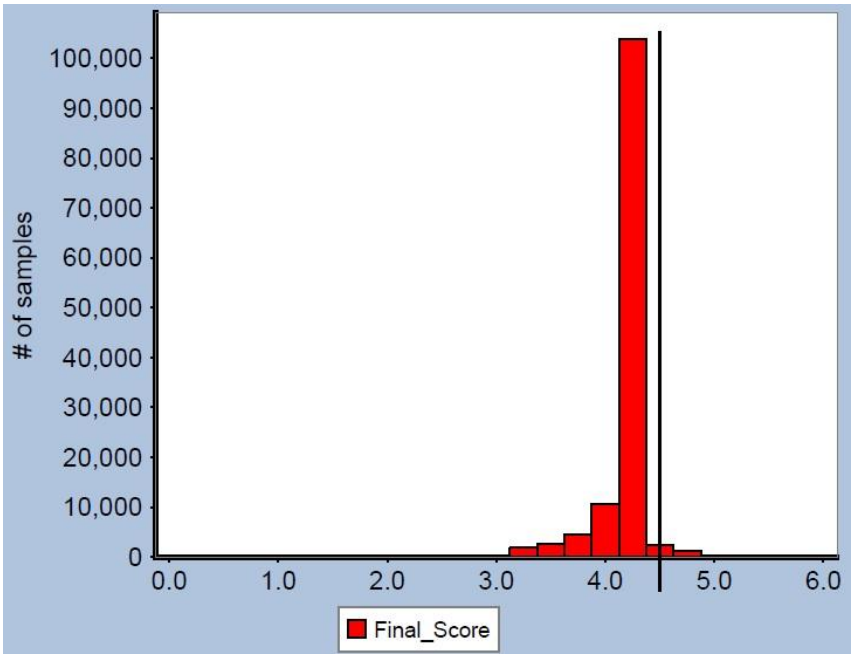


Fig. 3.8 Histogram representing the distribution of final score values across the 127,725 compounds remaining after filtering, with a line intersecting the cut-off value of 4.5.

3.3 Diversity Analysis & Compound Selection

With the dataset reduced to 2,538 compounds, several had to be selected for purchase and subsequent testing. Diversity analysis was therefore performed in order to best sample the chemical diversity of the molecules available. This allowed for as much chemical space to be covered as possible, given the resources/funds available. Four approaches were used to perform diversity analysis and compound selection.

Firstly the fifty top scoring compounds were selected, as not only had these clearly demonstrated good physicochemical properties, but they had also been strongly supported across the LBVS research. These would therefore be the most promising candidates if the LBVS results were considered alone. However, though these fifty compounds had the most support, it is unlikely that they fully represented the chemical diversity of the entire dataset. As research was concerned with finding novel lead like structures, it was essential that diversity analysis be performed in

order to select compounds for purchase and testing which represented the entire spread of chemical space sampled.

Following this the fifty most diverse compounds from the filtered dataset were selected based on their FCFP_4 fingerprints. A sample size of fifty was chosen as this allowed for a more focused selection amenable to visual inspection. Selecting a structurally diverse subset of compounds from the dataset allowed for a greater exploration of the activity space, and is particularly useful when little is known about the range of compounds active against a particular therapeutic target.[47] Fifty diverse structures according to their FCFP_4 fingerprints were selected using the *'Diverse Molecules'* component in Pipeline Pilot Student Edition v6.1.[22] Though the final scores of these compounds varied (Max=4.822; Min=4.5; Average=4.588), this method represented one way of sampling the range of chemical structures available. Diverse compound selection was performed according to the *'Diverse Compound Selection Protocol'* as described in the Experimental Chapter.

3.3.1 Clustering Analysis

The next two diversity methods involved the use of clustering analysis. Clustering analysis aims to divide a group of objects into clusters so that the objects within each cluster are similar, with objects taken from different clusters therefore being structurally dissimilar.[45] When molecules are clustered, a representative subset of one of more compounds can be selected from within that cluster, in order to represent that class of compounds, as illustrated by figure 3.9. Clustering analysis has found many uses, including in medicine and information sciences, and there are a large number of algorithms available for its implementation. The efficiency of

these algorithms various widely, with certain methods better suited to pharmaceutical applications and clustering databases of chemical structures.[47-50]

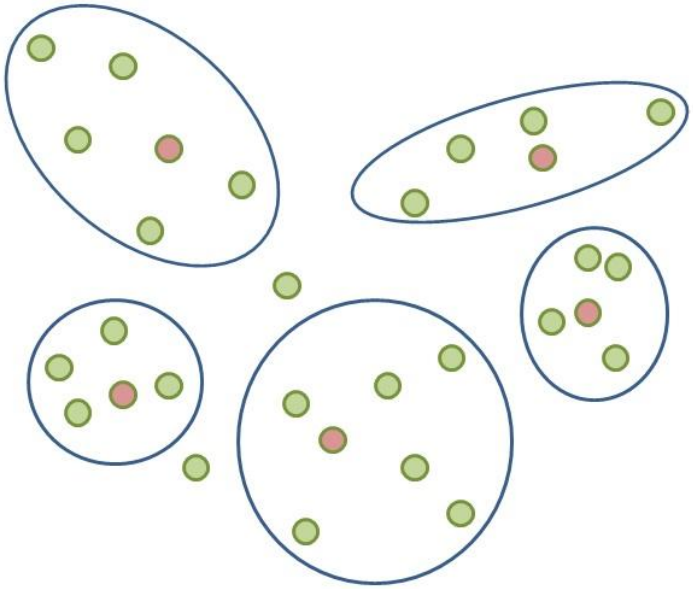


Fig. 3.9 Illustration of clustering analysis and the grouping together of similar compounds. Red circle show the selection of a representative molecule from a particular cluster.

The overall process of cluster based compound selection is as follows:

1. Generate descriptors for each compound in the dataset.
2. Calculate the similarity or distance between all compounds in the dataset.
3. Use a clustering algorithm to group the compounds within the dataset.
4. Select a representative subset by selecting one or more compounds from each cluster.

For step 1, the descriptors may include property values such as biological activity, or even topological indexes and structural fragments.[51, 52] For the two clustering analysis methods used here, FCFP_4 fingerprints were used to generate the descriptor values for each of the compounds in the dataset. Similarity measures may also be used in step 2 to quantify the degree of structural resemblance between pairs of molecules,[53] with Tanimoto once again used to calculate the similarity between

the compounds for both of the clustering methods employed. Where the two clustering methods employed vary is in step 3. Most clustering methods are non-overlapping, with each object belonging to just one cluster, (in overlapping methods compounds can belong to more than one cluster). The non-overlapping methods are divided into two classes, hierarchical and non-hierarchical. For step 3, one hierarchical method was used, and one non-hierarchical method.

3.3.1.1 Hierarchical Clustering

Hierarchical clustering methods organise compounds into clusters of increasing size, with small clusters of related compounds being grouped together into larger clusters. At one extreme each compound is in its own cluster, but after progressive joining of these smaller clusters, the compounds ultimately reside within a single cluster at the opposite extreme.[54] The successive levels and relationships between clusters can be visualised using a dendrogram, an example of which is in figure 3.10.

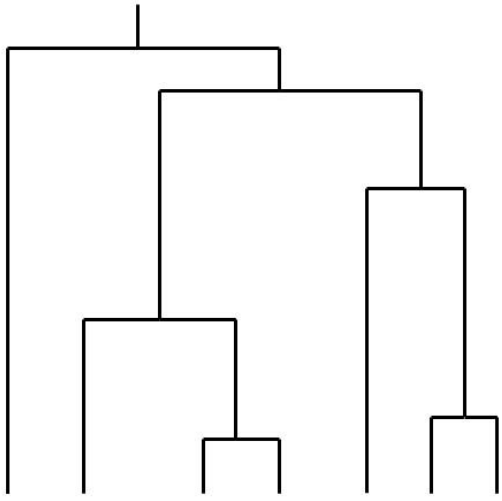


Fig. 3.10 A dendrogram representing a hierarchical clustering of seven compounds.

The dataset is analysed in an iterative manner, such that at each step either a pair of clusters are merged, or a single cluster is divided. Each level of the hierarchy represents a partitioning of the dataset. If a hierarchical method starts with all

compounds as singletons, that are then merged iteratively until all compounds are in a single cluster, the method is said to be agglomerative, that is from the bottom up in terms of the dendrogram.

Clusters are formed to minimise the total variance of the dataset.[55] The variance of a cluster is measured as the sum of the squared deviation from the mean of the cluster. For a cluster, i, of N_i objects where each object j is represented by a vector $r_{i,j}$ the mean (or centroid) of the cluster, $\bar{r}_i$ is given by equation 3.3, with the intracluster variance, E_i given by equation 3.4. The total variance is calculated as the sum of the intracluster variances for each cluster. At each iteration a pair of clusters is chosen whose merger leads to the minimum change in total variance. This is known as Ward's method.[55]

$$\bar{r}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} r_{i,j}$$

Eq. 3.3 Definition of the cluster centroid.

$$E_i = \sum_{j=1}^{N_i} \left(\left|r_{i,j} - \bar{r}_i\right|\right)^2$$

Eq. 3.4 Definition of intracluster variance.

The most commonly implemented hierarchical clustering methods are those belonging to the family of sequential agglomerative hierarchical non-overlapping (SAHN) methods. One particular example is AGNES,[56] or agglomerative nesting. SAHN methods are traditionally implemented using the stored matrix algorithm. Each cluster initially corresponds to an individual item, and as clustering proceeds, pairs of clusters are merged together and the number of clusters decreases by one. Eventually these evolves into just one cluster containing all items. The stored matrix algorithm is as follows:

1. Calculate the initial proximity matrix containing the pairwise proximities between all pairs of clusters (singletons) in the dataset.
2. Scan the matrix to find the most similar pair of clusters, and merge them into a new cluster (thus replacing the original pair).
3. Update the proximity matrix by inactivating one set of entries of the original pair and updating the other set (now representing the new cluster).
4. Repeat steps 2 and 3 until just one cluster remains.

Contrary to agglomerative methods, there are the divisive hierarchical clustering algorithms. These start with all compounds in a single cluster, and iteratively partitions one cluster into two (top to bottom on the dendrogram) until all compounds are singletons. This method is of particular use when only a small number of clusters is desired, so that only the first part of the hierarchy needs to be produced. Thus divisive methods can be faster than their agglomerative counterparts, though their overall performance is generally inferior.[57] This has been attributed to the fact that the initial criterion for partitioning a cluster is based on only a single descriptor, or is monothetic, unlike agglomerative methods which are polythetic.

When using hierarchical clustering methods it is necessary to choose a level from the hierarchy in order to define the appropriate number of clusters to represent the dataset. This corresponds to drawing an imaginary line across the dendrogram, with the number of vertical lines which intersect this line being equal to the number of clusters. This can be observed in figure 3.11, where the red line dissects the dendrogram, thus representing the seven molecules in four clusters. Visual inspections is a useful way to select the appropriate number of clusters, as the

dendrogram will accurately report the number of molecules in each cluster, allowing for a fair selection of clusters to best represent the dataset.

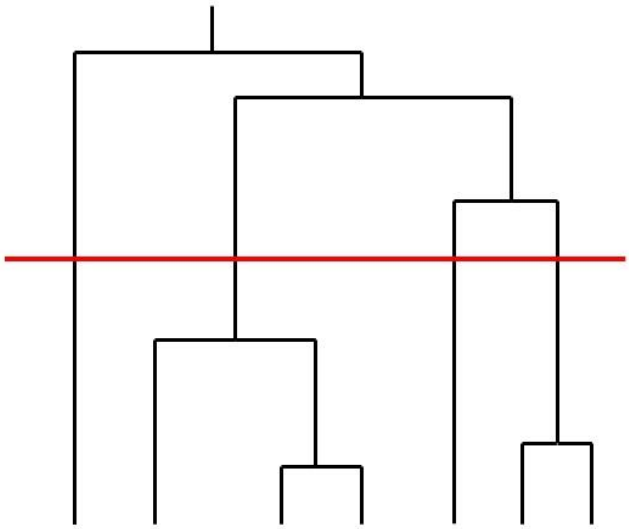


Fig. 3.11 Choosing the appropriate number of clusters to represent a dataset from a dendrogram. In this case four clusters represent the seven molecules.

The agglomerative hierarchical method AGNES[56] was used to perform clustering analysis on the 2,538 compounds. Clustering analysis was performed using the programming language R,[58] which is widely used for statistical software development and data analysis. R was utilised through Pipeline Pilot Student Edition v6.1.[22] AGNES is a hierarchical method which builds clusters from the bottom up, starting with all compounds as singletons. The developed protocol merged the molecules into clusters using Ward's method, with Euclidean distances acting as the distance matrix. In order to decide upon the optimum number of clusters the clustering analysis was first allowed to go to completion, so that all compounds could be represented by a single cluster. The resulting dendrogram was then inspected to decide upon the most appropriate number of clusters to use. It was required that the number of compounds selected be similar to the other methods, so inspection of the dendrogram began at around fifty clusters. This dendrogram is shown in figure 3.12. It was important that no cluster be more drastically

overpopulated than any other. As can be seen, if 52 clusters were selected, this allowed for one cluster to represent over a third of the entire dataset, illustrated by the red circle. To avoid this, a higher number of clusters were chosen. Investigation suggested that sixty four clusters was the minimum number required to give equally populated clusters.

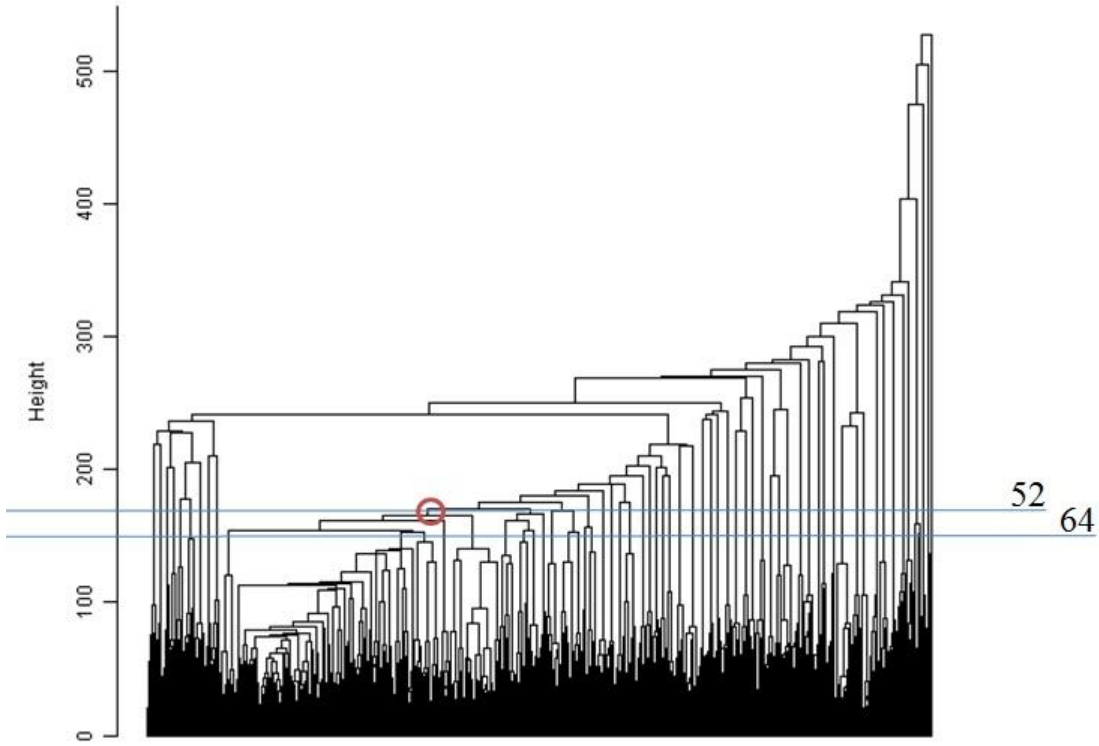


Fig. 3.12 Dendrogram representing the clustering of the 2,538 filtered compounds using AGNES. Blue lines dissecting the dendrogram represent the stated number of clusters. Red circle represents the potential for one hugely overpopulated cluster.

With the optimum number of clusters selected the analysis was repeated and each molecule assigned to its appropriate cluster. The most representative compound from each of the clusters was selected using FCFP_4 fingerprints, resulting in a subset of sixty four compounds which best represented the diversity of the dataset according to AGNES hierarchical clustering. AGNES clustering was performed according to the *'AGNES Clustering Protocol'* described in the Experimental Chapter.

3.3.1.2 Non-hierarchical Clustering

Non-hierarchical methods place compounds into clusters without forming a hierarchical relationship between them, with several non-hierarchical methods finding use in chemical applications such as the single-pass, relocation and nearest neighbour methods.[47, 50] Single-pass methods cluster objects based upon only a single pass through the dataset. The first compound encountered is assigned to the first cluster, with the next compound also assigned to that cluster should its similarity exceed some specified threshold value. Otherwise it is assigned to a new cluster. This process is repeated until all compounds have been assigned to clusters. Though this method is very fast and simple to implement, its major drawback is that it is order dependent. That is, if the compounds are rearranged and scanned in a different order, then the resulting clusters may also be different.

The best known relocation method is the K-means method, of which there exist many variants and different algorithms for its implementation.[59] The basic algorithm acts to minimise the sum of the squared Euclidean distances between each item in a cluster and the cluster centroid. The first step is to choose a set of c seed compounds which are usually selected at random. The remaining compounds are then assigned to the nearest seed to give an initial set of c clusters. The centroids of the clusters are then calculated and the objects are reassigned (or relocated) to the nearest cluster centroid. This process of calculating cluster centroids and relocating the compounds is repeated until no objects change clusters, or until a user defined number of iterations has been performed. This method is dependent upon the initial set of cluster centroids, and different results will usually be found for different initial seeds.[60] Because of this, relocation methods can be adversely affected by outlier

compounds which will appear as singletons, as they are not sufficiently similar to anything else. Nevertheless, the computational efficiency and mathematical foundation of relocation methods have made them very popular.

The relocation method CLARA[56] was employed as an additional means to assess the diversity of the 2,538 filtered compounds. CLARA (clustering large applications) is particularly useful for sampling large datasets, similar to the dataset which was in question here. CLARA diversity analysis was performed using the R[58] programming language in Pipeline Pilot Student Edition v6.1.[22] Ward's method was used together with Euclidean distance as the distance matrix. Unlike AGNES, it was not necessary to inspect a dendrogram in order to select the most appropriate number of clusters, but to instead simply define the number of clusters that were required. Fifty clusters were chosen, with the central molecule from each selected based upon its positioning in the cluster according to its Euclidean distance. This identified a further 50 compounds from diversity analysis. CLARA clustering was performed according to the *'CLARA Clustering Protocol'* as described in the Experimental Chapter.

3.4 Visual Inspection

The four methods of diversity analysis resulted in four subsets of compounds (highest scoring; fingerprint analysis; AGNES clustering; CLARA clustering) which each sampled the chemical space of the 2,538 filtered compounds differently. These subsets were merged and of the 214 entries recorded, only 205 were unique, as a number of the compounds appeared across multiple methods (one compound appeared in three diversity methods; seven compounds appeared in two diversity methods). Lists containing the 205 chemical structures were distributed amongst members of the antimalarial drug discovery group at Liverpool University. They

were visually inspected so that potentially undesirable molecules could be spotted prior to final selection. This was to draw upon the extensive experience of senior members within the group as to the synthesisability of the chemotypes, and provided an ideal opportunity to spot any compounds that had slipped through despite rigorous filtering.

In total 66 additional compounds were removed, resulting in only 139 remaining. There were several reasons as to why certain compounds were recommended for removal. Several appeared to be very large, and given that chemical optimisation generally increases a lead compounds MW, their removal seemed wise. Some were simply too simple, consisting of very basic, small ring systems, whilst others closely resembled chemotypes that had already been investigated within the group, such as quinolone and chromone. Perhaps the most crucial observation was the removal of all compounds that contained a furan ring. This observation alone highlighted the importance of visual inspection, as it was initially thought that this toxicophore would have been included within the Rishton fragments. It was proposed that the furan moiety may potentially react with nucleophiles in the body under acid conditions, a possible mechanism for which is illustrated in figure 3.13.

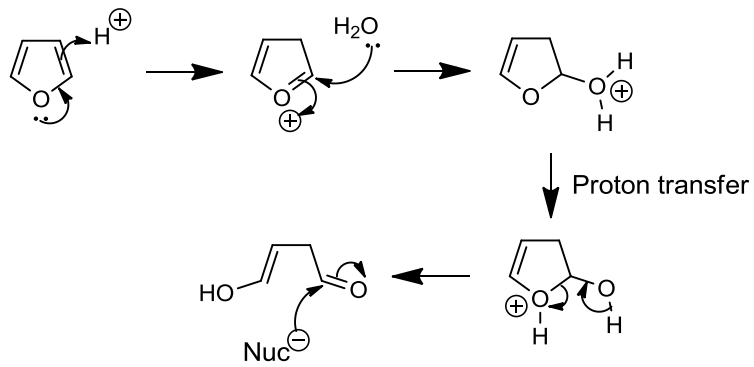


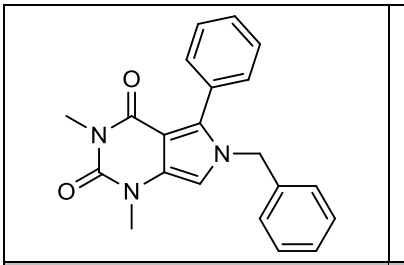
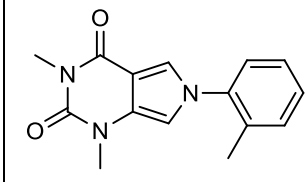
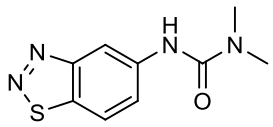
Fig. 3.13 Mechanism to illustrate potential toxicity concerns with regard to furan.

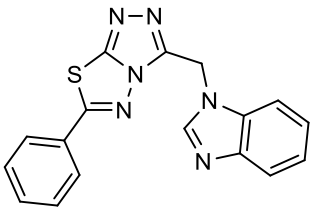
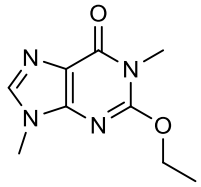
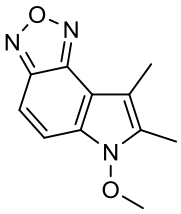
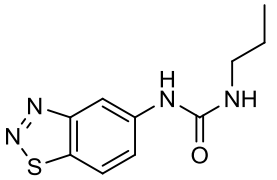
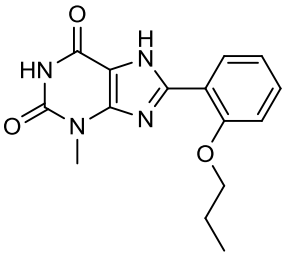
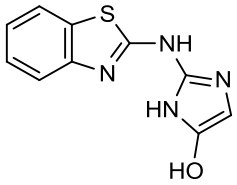
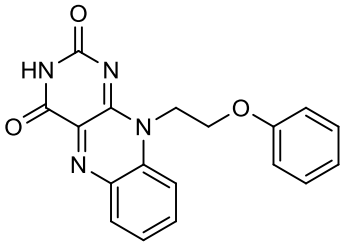
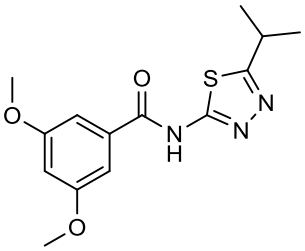
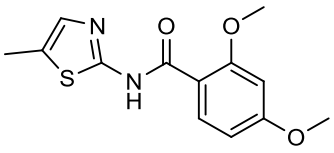
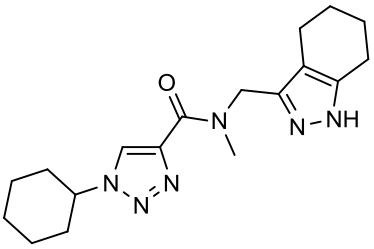
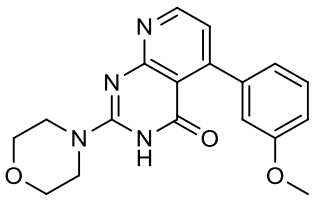
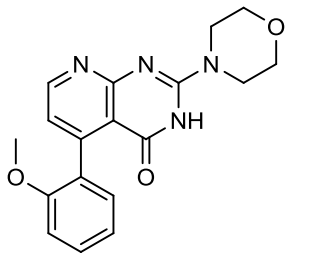
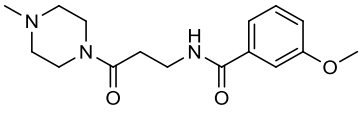
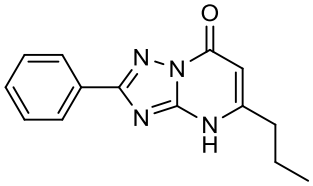
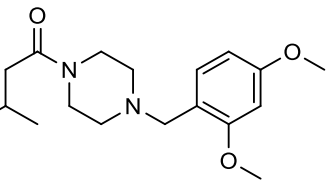
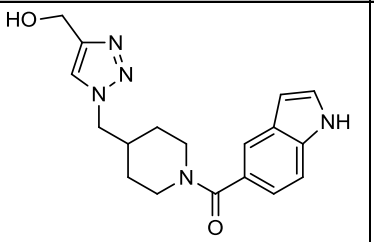
3.5 Final Selection/Purchasing Compounds

The final subset of 139 compounds were deemed to be the most promising lead like candidates to result from the extensive LBVS, filtering and diversity analysis work with regard to the *Pf*bc$_1$ biochemical target (see appendix for all 139 compounds). It was these compounds, or a further subset thereof, which were therefore required for biological testing. Using the ZINC website[61] it is possible to generate vendor information for a set of compounds, by simply inputting the necessary ZINC IDs. This was done for the 139 compounds, but unfortunately, during the course of this research the ZINC lead like library was updated from version 7 (2,710,002 compounds) to version 8 (4,390,615 compounds). Following this update, many of the compounds which had previously been reported as available in ZINC version 7, were no longer available. Target synthesis of all the compounds was far too costly, thus a much smaller subset of the compounds was settled upon.

The supplier ChemBridge[62] was able to offer 19 of the compounds from our list of 139 structures, each at quantities of 5 milligrams (mg) which was sufficient for biological analysis. The structures and labels for these purchased compounds are shown in table 3.5.

Table. 3.5 Compounds purchased from LBVS, together with their Final_Score (in brackets).

		
VS01 (4.931)	**VS02** (4.5)	**VS03** (4.625)

		
VS04 (4.625)	**VS05** (4.625)	**VS06** (4.5)
		
VS07 (4.625)	**VS08** (4.5)	**VS09** (4.5)
		
VS10 (4.568)	**VS11** (4.739)	**VS12** (4.661)
		
VS13 (4.625)	**VS14** (5.038)	**VS15** (4.980)
		
VS16 (4.884)	**VS17** (4.5)	**VS18** (4.893)
		
VS19 (4.625)		

With a high scoring and structurally diverse selection of compounds purchased, biological testing could be performed to evaluate the potential of these structures as

novel lead like chemotypes, active against *Pf*bc$_1$. Chapter IV will discuss the details of the biological testing that was performed, together with analysis of the resulting active chemotypes.

3.6 References

1. J. J. Irwin and B. K. Shoichet, *Journal of Chemical Information and Modeling*, 2005, **45**, 177-182.
2. S. J. Teague, A. M. Davis, P. D. Leeson and T. Oprea, *Angew. Chem.-Int. Edit.*, 1999, **38**, 3743-3748.
3. F. Miklos, *Drug Discovery Today*, 2006, **11**, 421-428.
4. M. Stahl and M. Rarey, *Journal of Medicinal Chemistry*, 2001, **44**, 1035-1042.
5. S. S. So, S. P. van Helden, V. J. van Geerestein and M. Karplus, *Journal of Chemical Information and Computer Sciences*, 2000, **40**, 762-772.
6. C. Bissantz, G. Folkers and D. Rognan, *Journal of Medicinal Chemistry*, 2000, **43**, 4759-4767.
7. S. Ha, R. Andreani, A. Robbins and I. Muegge, *Journal of Computer-Aided Molecular Design*, 2000, **14**, 435-448.
8. S. Zhong, Y. Zhang and Z. Xiu, *Current Opinion in Drug Discovery and Development*, 2010, **13**, 326-334.
9. P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *Journal of Medicinal Chemistry*, 1999, **42**, 5100-5109.
10. G. E. Terp, B. N. Johansen, I. T. Christensen and F. S. Jorgensen, *Journal of Medicinal Chemistry*, 2001, **44**, 2333-2343.
11. T. Cheng, X. Li, Y. Li, Z. Liu and R. Wang, *Journal of Chemical Information and Modeling*, 2009, **49**, 1079-1093.
12. R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake and J. B. Matthew, *J. Mol. Graph.*, 2002, **20**, 281-295.
13. R. Rajamani and A. C. Good, *Current Opinion in Drug Discovery and Development*, 2007, **10**, 308-315.
14. J. C. Baber, W. A. Shirley, Y. Gao and M. Feher, *Journal of Chemical Information and Modeling*, 2005, **46**, 277-288.
15. H. Gohlke and G. Klebe, *Curr. Opin. Struct. Biol.*, 2001, **11**, 231-235.
16. N. Paul and D. Rognan, *Proteins*, 2002, **47**, 521-533.
17. C. M. R. Ginn, P. Willett and J. Bradshaw, *Perspect. Drug Discov. Design*, 2000, **20**, 1-16.
18. N. Salim, J. Holliday and P. Willett, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 435-442.
19. J. W. Godden, J. R. Furr, L. Xue, F. L. Stahura and J. Bajorath, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 21-29.
20. N. Baurin, J. C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot and L. Morin-Allory, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 276-285.
21. J. W. Raymond, M. Jalaie and M. P. Bradley, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 601-609.
22. SciTegic, *Pipeline Pilot Student Edition v6.1*, Accelrys, Inc, San Diego, CA, 2007.
23. W. L. Jorgensen and E. M. Duffy, *Adv. Drug Deliv. Rev.*, 2002, **54**, 355-366.
24. C. A. Lipinski, *Drug Discovery Today: Technologies*, 2004, **1**, 337-341.
25. W. L. Jorgensen, *Science*, 2004, **303**, 1813-1818.
26. E. H. Kerns and L. Di, *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization*, Elsevier, 2008.
27. R. Gilbert M, *Drug Discovery Today*, 2003, **8**, 86-96.
28. M. S. Lajiness, M. Vieth and J. Erickson, *Current Opinion in Drug Discovery and Development*, 2004, **7**, 470-477.
29. R. Sharma, A. S. Lawrenson, N. E. Fisher, A. J. Warman, A. E. Shone, A. Hill, A. Mbekeani, C. Pidathala, R. K. Amewu, S. Leung, P. Gibbons, D. W. Hong, P. Stocks, G. L. Nixon, J. Chadwick, J. Shearer, I. Gowers, D. Cronk, S. P. Parel, P. M. O'Neill, S. A. Ward, G. A. Biagini and N. G. Berry, *Journal of Medicinal Chemistry*, 2012, **55**, 3144-3154.
30. C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235-249.
31. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 2001, **46**, 3-26.
32. D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *Journal of Medicinal Chemistry*, 2002, **45**, 2615-2623.
33. P. Schlagenhauf, *Journal of Travel Medicine*, 1999, **6**, 122-133.
34. A. Fletcher and R. Shepherd, *Use of (+)-mefloquine for the treatment of malaria*, 2002.

35. G.-Q. Lin, Q.-D. You and J.-F. Cheng, *Chiral Drugs: Chemistry and Biological Action*, Wiley, 2011.
36. G. M. Rishton, *Drug Discovery Today*, 1997, **2**, 382-384.
37. G. M. Rishton, *Drug Discovery Today*, 2003, **8**, 86-96.
38. D. P. Williams and D. J. Naisbitt, *Current Opinion in Drug Discovery & Development*, 2002, **5**, 104-115.
39. R. Brenk, A. Schipani, D. James, A. Krasowski, I. H. Gilbert, J. Frearson and P. G. Wyatt, *ChemMedChem*, 2008, **3**, 435-444.
40. D. P. Williams, *Toxicology*, 2006, **226**, 3-13.
41. K. Park, D. P. Williams, D. J. Naisbitt, N. R. Kitteringham and M. Pirmohamed, *Toxicol Appl Pharmacol*, 2005, **207**, 425-434.
42. N. Baurin, R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parratt, P. Greaney, D. Morley and R. E. Hubbard, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 643-651.
43. M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller and N. Ramsden, *Journal of Chemical Information and Computer Sciences*, 1999, **39**, 897-902.
44. W. P. Walters and M. A. Murcko, in *Virtual Screening for Bioactive Molecules*, Wiley-VCH Verlag GmbH, Editon edn., 2008, pp. 15-32.
45. D. J. Huggins, A. R. Venkitaraman and D. R. Spring, *ACS Chem. Biol.*, 2011, **6**, 208-217.
46. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell (4th ed.)*, 2002.
47. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
48. P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth, UK, 1987.
49. J. M. Barnard and G. M. Downs, *Journal of Chemical Information and Computer Sciences*, 1992, **32**, 644-649.
50. G. M. Downs and J. M. Barnard, in *Reviews in Computational Chemistry*, John Wiley & Sons, Inc., Editon edn., 2003, pp. 1-40.
51. R. D. Brown and Y. C. Martin, *Journal of Chemical Information and Computer Sciences*, 1996, **36**, 572-584.
52. R. D. Brown and Y. C. Martin, *Journal of Chemical Information and Computer Sciences*, 1997, **37**, 1-9.
53. P. Willett, J. M. Barnard and G. M. Downs, *Journal of Chemical Information and Computer Sciences*, 1998, **38**, 983-996.
54. F. Murtagh, *Comput. J.*, 1983, **26**, 354-359.
55. J. H. Ward, Jr., *Journal of the American Statistical Association*, 1963, **58**, 236-244.
56. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
57. V. Rubin and P. Willett, *Analytica Chimica Acta*, 1983, **151**, 161-166.
58. *R version 2.9.0*, The R Foundation for Statistical Computing, 2009.
59. E. W. Forgy, *Biometrics*, 1965, **21**, 768-&.
60. G. W. Milligan, *Psychometrika*, 1980, **45**, 325-342.
61. ZINC, *http://zinc.docking.org/*.
62. ChemBridge, *http://www.chembridge.com/*.

Chapter IV

Ligand Based Virtual Screening Testing & Analysis

4. Ligand Based Virtual Screening Testing & Analysis 199

4.1 Biological Testing 199

4.1.1 Whole Cell Growth Inhibition Bioassay 201

4.1.2 Complex I (NDH2) Bioassay 203

4.1.3 Bovine Complex III (bc_1) Bioassay 205

4.2 Analysis of Active Compounds 207

4.2.1 Ligand Efficiency 209

4.2.2 Novelty of the Chemotypes 210

4.3 VS01 213

4.4 VS09 215

4.5 VS10 219

4.6 VS16 and VS18 223

4.7 Summary of Testing/Analysis 224

4.8 References 226

4. Ligand Based Virtual Screening Testing & Analysis

This chapter is concerned with the third phase of the molecular design loop (fig. 4.1), biological testing. If the synthesis of the nineteen compounds described in table 2.1 and their testing against *Pf*bc$_1$ is considered to be one complete iteration of the loop, then the use of their structures in Chapter II represents the beginnings of a new iteration of the loop through LBVS. Phase two (chemical synthesis) of this iteration represents the purchasing of the compounds from the third party vendor (Chapter III), with the second iteration of the loop finally completed as follows.

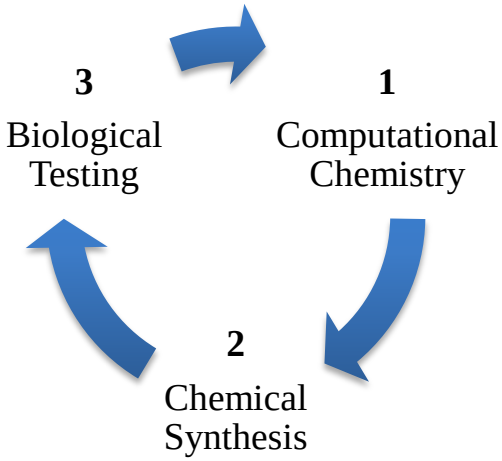


Fig. 4.1 Molecular design loop.

4.1 Biological Testing

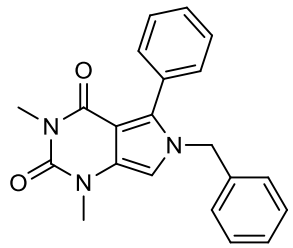
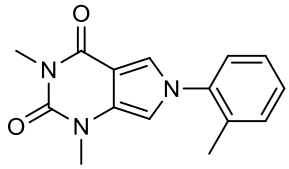
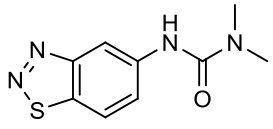
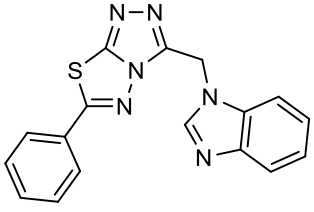
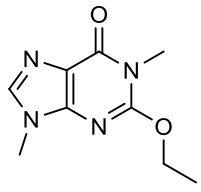
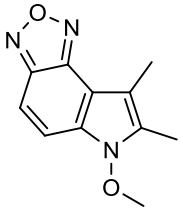
Biological testing is crucial to all aspects of the drug discovery process,[1] allowing compounds to be assigned quantitative and/or qualitative information with regard to their pharmacological activity. Bioassays offer an excellent way to measure this *in vitro*, and are particularly useful in the early stages of drug development. The use of bioassays in antimalarial drug design is well documented, with research continually ongoing to optimise, develop and refine the techniques,[2-5] as their reliability and reproducibility are key.[6] *In vitro* testing is usually fairly cheap and straight forward

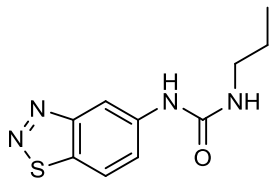
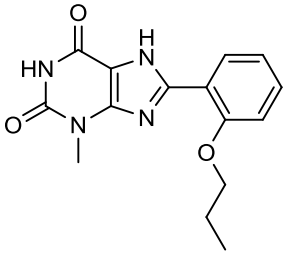
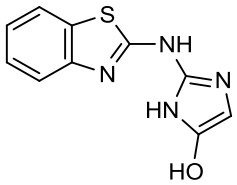
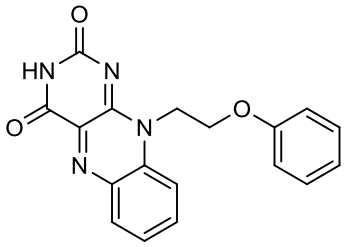
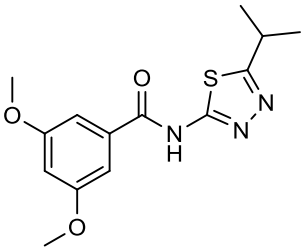
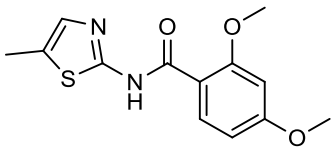
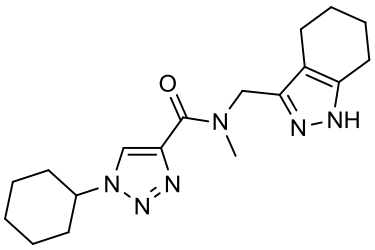
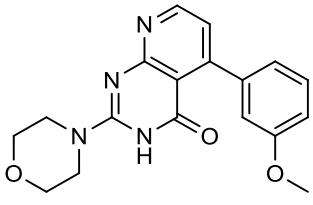
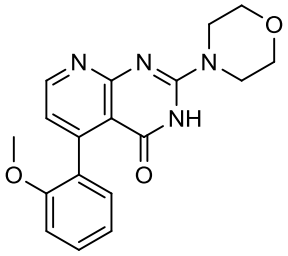
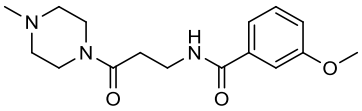
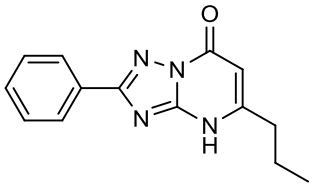
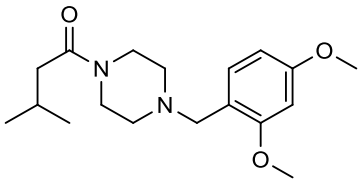
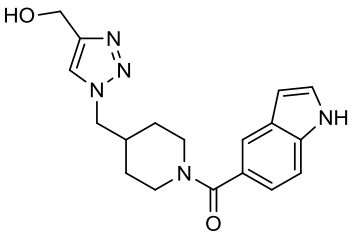
to perform, and may often be automated such that large numbers of compounds can be tested, making them suited to high throughput screening.[7] Promising drug candidates may ultimately undergo further *in vitro* testing, and eventual *in vivo* analysis.

A number of bioassays were employed in order to test the antimalarial potential of the 19 compounds purchased as a result of LBVS. Selectivity is crucial in drug discovery to avoid off target toxicity and other undesirable effects. Once a compound which inhibits growth of the malaria parasite had been found, further testing could be performed to allow assessment of its selectivity. Ideally though, an antimalarial compound would be selective for the malaria parasite, having little or no ill effect on the patient.

It is common practice within antimalarial drug design to first identify compounds active against the malaria parasite using whole cell screening, and to then screen the actives with specific bioassays to determine possible modes of action.[8] Therefore the 19 purchased compounds, as shown in table 4.1, were tested against several bioassays.

Table. 4.1 Compounds purchased from LBVS, together with their Final_Score (in brackets).

		
VS01 (4.931)	**VS02** (4.5)	**VS03** (4.625)
		
VS04 (4.625)	**VS05** (4.625)	**VS06** (4.5)

		
VS07 (4.625)	**VS08** (4.5)	**VS09** (4.5)
		
VS10 (4.568)	**VS11** (4.739)	**VS12** (4.661)
		
VS13 (4.625)	**VS14** (5.038)	**VS15** (4.980)
		
VS16 (4.884)	**VS17** (4.5)	**VS18** (4.893)
		
VS19 (4.625)		

4.1.1 Whole Cell Growth Inhibition Bioassay

The all important first step during biological testing was to determine whether any of the 19 compounds showed potential as antimalarials. This required the use of a whole cell growth inhibition assay, which was performed according to the *'Whole*

Cell Growth Inhibition Assay (3D7) Protocol' as described in the Experimental Chapter.[2, 9-11] The assay was performed using the CQS strain of *P. falciparum*, 3D7. IC_{50} values were recorded for each of the compounds that inhibited the parasite, and though the assay did not consider compound selectivity, it provided an initial means for assessing a compounds antimalarial potential.

Compounds were first tested to a maximum *in vitro* concentration of 1.0 µM. However, when this failed to produce any quantitative hits, the maximum concentration for testing was increased to 10 µM. Of the 19 compounds, 5 were subsequently found to have IC_{50} values less than 10 µM, ranging between 4.53 and 8.18 µM. These results are shown in table 4.2. These activity values were highly encouraging, with similar work involving the use of LBVS to discover novel plasmepsin inhibitors (a family of malarial parasitic aspartyl proteases), reporting whole cell growth inhibition values of as much as 40 µM.[9] There are also several other instances where µM hits have led to nM lead structures.[10-12]

Table. 4.2 *In vitro* IC_{50} values against 3D7 strain of *P. falciparum* for the LBVS hits.

ID	IC_{50} (µM) 3D7
VS01	5.24
VS09	4.53 ± 1.86
VS10	6.41 ± 1.73
VS16	5.39
VS18	8.18

All compounds were tested in triplicate. Three reported activity values only once, whilst the other two (VS09 and VS10) were consistently reproducible (shown by the presence of SD values). As the goal was to identify novel antimalarial chemotypes amenable to further optimisation, these initial activity values looked very encouraging. Recent work involving the use of virtual screening to identify novel and selective pteridine reductase 1 (PTR1) inhibitors for the treatment of human African trypanosomiasis (HAT), or sleeping sickness, tested the selected compounds

at a concentration of 100 μM.[13] Though high, it was rationalised that given the low MW of the screened hits, and that the intention was to optimise these structures further, inhibition at 100 μM was deemed encouraging. In this instance, the two most potent hits which observed just over 50% inhibition at 100 μM, led to PTR1 inhibitors with low nM potency, and favourable physicochemical properties. The same approach may therefore be possible for the 5 active compounds described in table 4.2.

Given that nineteen compounds were purchased and five reported some *in vitro* activity against 3D7, the hit/success rate for the LBVS work described in Chapter II was therefore over 26%. This is a crude estimate, given that the compounds were partly selected owing to their availability, but this figure is certainly significant, and higher than what would have been achieved had the compounds simply been selected at random (~0.1% hit rate)[14-16] from the 2.7 million in ZINC.[17]

For the 5 active compounds, testing could begin with regard to assessing their selectivity and toxicity. Unfortunately, owing to the difficulties and expense of isolating the *Pf*bc_1 enzyme, testing was not performed against this bioassay. However, two other assays were employed that would each provide additional information about the 5 active compounds.

4.1.2 Complex I (NDH2) Bioassay

*Pf*NDH2 is an alternative target for antimalarial drug design (Chapter I),[18] thus the NDH2 assay was performed to see whether any of the compounds in table 4.2 were active against this enzyme. This was useful in assessing the potential cross over between bc_1 and NDH2 activity.[19, 20] The NADH:ubiquinone oxidoreductase enzyme (complex I) catalyses the reaction shown in figure 4.2, and it is this reaction

which can be monitored *in vitro* to assess the effect of a particular compound against complex I.[5]

$$NADH + Quinone \rightarrow NAD^{+} + Quinol$$

Fig. 4.2 Reaction taking place during the NDH2 bioassay. Process catalysed by complex I.

The NDH2 bioassay is used to monitor the enzyme inhibition of a query compound, using the *'Complex I (NDH2) Bioassay Protocol'* as described in the Experimental Chapter.[5] The *in vitro* reaction is monitored spectrophotometrically at 283 and 340 nm, with these values monitoring the quinone reduction and NADH oxidation respectively. For each compound tested two inhibition values were reported, one at 283 nm and the other at 340 nm. Though the two are generally similar in magnitude, previous work found that monitoring the reaction at 340 nm, (that is NADH oxidation and not quinone reduction) reduced the potential interference from inhibitors and generated more robust assay performance measures.[5]

The results of the NDH2 bioassay for the 5 hits are reported in table 4.3. The compounds were each tested at a concentration of 28 μM, which was the maximum drug concentration the assay could deal with before secondary effects were observed, leading to inaccurate results. Percentage inhibition values at this concentration were recorded.

Table. 4.3 *In vitro* percentage inhibition of NDH2 at 340 and 283 nm.

	Inhibition of NDH2 at 28 μM (%)	
ID	**340 nm**	**283 nm**
VS01	0.8	2.6
VS09	5.9	5.0
VS10	3.3	4.1
VS16	4.6	7.6
VS18	6.0	4.1

As can be seen, all 5 of the compounds produced little or no inhibition of NADH activity, with VS18 giving the highest percentage inhibition of only 6% (at 340nm).

The NDH2 assay was performed to see whether the compounds were killing the parasite through complex I inhibition, but this is clearly not the case. This offers indirect support for the selectivity of the compounds as *Pf*bc_1 inhibitors. However, had the compounds inhibited NDH2 this would still have been a positive result, as perhaps more than one target was being hit (dual inhibition), and so resistance would be more difficult to acquire owing to multiple sites of action. From these results it was therefore possible to conclude that the 5 hits still looked to show much potential, particularly VS09 and VS10, thanks to the reproducibility of their *in vitro* activities.

4.1.3 Bovine Complex III (bc_1) Bioassay

Owing to difficulties in obtaining enough material to perform the *Pf*bc_1 bioassay (with estimates being at least nine months to grow enough parasite), additional information relating to the hits was collected in preparation. The bovine bc_1 bioassay was performed according to the *'Bovine Complex III (bc_1) Bioassay Protocol'* detailed in the Experimental Chapter. This assay is very similar to the one which will be performed when sufficient amounts of the *Pf*bc_1 enzyme are available, only using the bovine bc_1 complex as oppose to that of *plasmodium.* The principal behind the assay is to measure the effect (if any) a compound has on cytochrome c reductase activity. As was discussed extensively in the introduction, during the Q-cycle, a proton gradient is generated across the mitochondrial membrane, leading to the release of four protons into the inter membrane space.[21, 22] During this process, cytochrome c becomes reduced whilst ubiquinol (QH2) is oxidised to ubiquinone (Q). The assay monitors this reduction of cytochrome c spectrophotometrically,[2] with compounds that halt reduction therefore acting as bc_1 inhibitors.

Though this assay measures bovine bc_1 inhibition, it has been found to also provide a means for identifying potentially toxic compounds. It has been observed that compounds which show strong inhibition of bovine bc_1, often produce cardiotoxicity in humans.[2, 3] This may be attributed to the high degree of sequence identity between the human and bovine binding sites.[23] An example of this is SMA, which though highly active towards *Pf*bc_1 (IC_{50} of 12 ± 1 nM), shows similar potency towards bovine bc_1 (IC_{50} of 2.4 nM), and thus strong inhibition of human liver bc_1 (IC_{50} of 15 ± 0.2 nM). Conversely, the potent dihydroacridinedione compound WR249685 (fig. 4.3) was found to be highly selective for *Pf*bc_1 (IC_{50} of 3 ± 2 nM), but showed poor inhibition of both bovine bc_1 (IC_{50} of > 13,800 nM) and human liver bc_1 (IC_{50} of > 13,800) nM). Given this trend, it is preferable that compounds avoid inhibiting bovine bc_1, and show improved selectivity for the parasite.

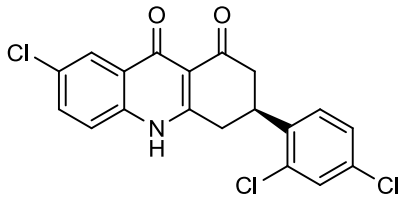


Fig. 4.3 WR 249685.

The 5 hits produced bovine bc_1 inhibition values as shown in table 4.4. Similar to the complex I bioassay, the assay had a maximum drug concentration it could deal with before secondary effects were observed, which in this case was 56 µM. Compounds were therefore tested up to this maximum concentration, and their percentage inhibitions recorded.

Table. 4.4 *In vitro* percentage inhibition of bovine bc_1.

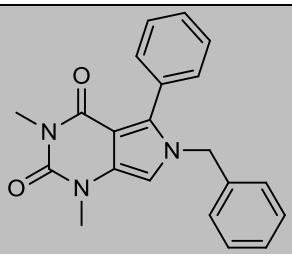
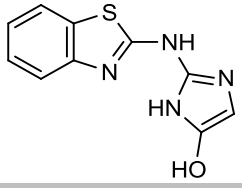
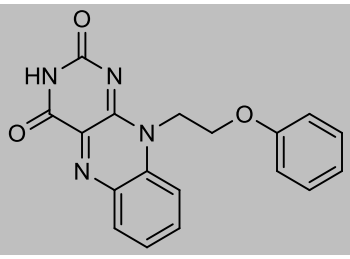
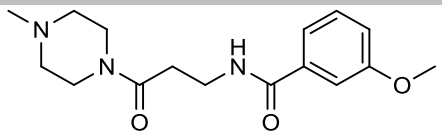
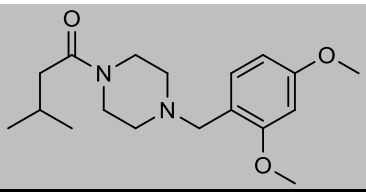
ID	Inhibition of bovine bc_1 at 56 µM (%)
VS01	0
VS09	1.1
VS10	3.4
VS16	9.6
VS18	0

The highest inhibition of the 5 compounds was found to be just under 10% (VS16), whilst the others all had very little or no inhibition. This is encouraging given the high concentration at which they were tested, and suggests that given previous observations,[2] these compounds may avoid the associated cardiotoxicity concerns of other bc_1 inhibitors. If indeed (after future testing) these compounds prove to inhibit *Pf*bc_1, then these results support their selectively and their avoidance of bovine bc_1 inhibition.

4.2 Analysis of Active Compounds

The results of biological testing are summarised in table 4.5, together with the chemical structures of the hits. Analysis of the actives involved careful examination of any trends in the data, as well as looking more thoroughly at the promising lead like candidates.

Table. 4.5 Summary of *in vitro* testing results.

Structure	ID	3D7 IC_{50} (μM)	NDH2 (% inhibition) 340 nm	NDH2 (% inhibition) 283 nm	Bovine bc_1 (% inhibition)
	VS01	5.24	0.8	2.6	0
	VS09	4.53 ± 1.86	5.9	5.0	1.1
	VS10	6.41 ± 1.73	3.3	4.1	3.4
	VS16	5.39	4.6	7.6	9.6
	VS18	8.18	6.0	4.1	0

To summarise, 5 of the 19 purchased compounds reported activity against the 3D7 strain of the malaria parasite. The reported IC_{50} values themselves are all encouraging, particularly with regard to recent precedent in the literature for novel hit to lead structures of antimalarials.[24-26] There were slight concerns with regard to the reproducibility of the *in vitro* activities, but for VS09 and VS10 this was not a concern, as both compounds performed consistently, reporting low SD values. Unfortunately however, only single IC_{50} values were recorded for VS01, VS16 and VS18. Therefore from 3D7 testing alone, VS09 and VS10 appeared to be the most

promising hits. Additionally, NDH2 inhibition was consistently low across all 5 compounds, as was bovine bc_1 inhibition. This is particularly significant as it reduces the number of possible sites of action for the compounds, and also indicates that they avoid hitting an enzyme known to be indicative of cardiotoxicity in humans.

4.2.1 Ligand Efficiency

Ligand efficiency (LE) is a measure of the binding energy between a ligand and its binding partner, such as a receptor or enzyme.[27] It is essentially a measure of the per atom potency of a compound, and is commonly used in drug discovery efforts to assist in narrowing down lead compounds with favourable physicochemical and pharmacological properties.[28] Ligand efficiency was initially defined numerically as the ratio of the Gibbs free energy (ΔG) to the number of non-hydrogen atoms (N) in a compound.[29] However, given that non-hydrogen atoms can be of many different types, and that MW is a key property in a compound, the concept of LE has been extended.[30] The binding efficiency index (BEI) offers a means to associate the potency of a compound to its MW, on a per kDa scale.[28] BEI is calculated using to equation 4.1, where $p(IC_{50})$ is defined as $-\log(IC_{50})$.

$$BEI = \frac{p(IC_{50})}{MW(kDa)}$$

Eq. 4.1 Binding efficiency index formula.

The BEI values for the 5 active compounds and CQ are reported in table 4.6. The inclusion of CQ enabled a comparison between the results, as it is a known, potent inhibitor of 3D7. A recent publication has shown CQ to have an IC_{50} value of around 15 nM.[31] A reference value for BEI of 27.0 is also reported in table 4.6 (note that this is a general value and not something specific to 3D7), based on idealised

values for MW and IC_{50}.[30] MW was taken to be 0.333 kDa, as this was found to be near the mean value of MW based on a large sample of marketed oral drugs,[32, 33] with the IC_{50} value simply defined as 1 nM.

Table. 4.6 Calculation of BEI for the LBVS hits.

Molecule	IC50 3D7 (μM)	IC50 3D7 (M)	$p(IC_{50})$	MW (Da)	MW (kDa)	BEI
CQ	0.015	0.000000015	7.82	319.9	0.3199	**24.5**
VS01	5.24	0.00000524	5.28	345.1	0.3451	**15.3**
VS09	4.53	0.00000453	5.34	232.0	0.2320	**23.0**
VS10	6.41	0.00000641	5.19	334.1	0.3341	**15.5**
VS16	5.39	0.00000539	5.27	305.2	0.3052	**17.3**
VS18	8.18	0.00000818	5.09	320.2	0.3202	**15.9**
Reference	0.001	0.000000001	9	333.0	0.3330	**27.0**

The BEI for CQ was found to be 24.5, very similar to the reference value (27.0). CQ could therefore be used as a reference point to which the other compounds can be compared for their 3D7 potency. In CQ, each atom clearly contributes highly to its overall activity, making it highly ligand efficient. Though the BEI values for the purchased hits are lower than that of CQ, they are still largely comparable. In particular, VS09 was found to have a highly promising BEI value of 23.0. Considering how small the molecule is (MW of 232), it undoubtedly shows much potential for subsequent SAR investigation and chemical optimisation.

4.2.2 Novelty of the Chemotypes

Several datasets of compounds have recently been published which report the structures and activity values of many compounds known to inhibit the malaria parasite. Two such datasets are the GSK and St Jude's databases.[34, 35] GSK screened over two million compounds and tested them against 3D7 *P. falciparum* at a concentration of 2 μM. Around 13,000 (13,519) of the compounds had greater than 80% inhibition, and were thus reported in the GSK database. Similarly, the St

Jude's database reported the identity of around 1,300 (1,236) hits from an initial screen of 300,000 compounds which observed greater than 80% inhibition of 3D7, at a concentration of 7 μM. These sets of compounds with known antimalarial behaviour were used to assess the novelty of the 5 active hits identified from LBVS.

There were a few core motifs which represented the central chemotypes of the 5 active hits (table 4.5). These were: pyrrolopyrimidine-2,4-dione; benzothiazole; isoalloxazine ring; piperazine. A substructure search of the four chemotypes in both the GSK and St Jude's databases was performed using the *'Substructure Searching Protocol'* as described in the Experimental Chapter. Table 4.7 illustrates the number of times a particular chemotype appeared in the database.

Table. 4.7 Frequency of chemotypes in the GSK and St Jude's databases.

	Frequency	
Chemotype	**GSK (13,519 compounds)**	**St Jude's (1,236 compounds)**
Pyrrolopyrimidine-2,4-dione	0	0
Benzothiazole	79	62
Isoalloxazine Ring	0	0
Piperazine	1208	114

As can be seen, the pyrrolopyrimidine-2,4-dione and isoalloxazine chemotypes were not observed in either of the databases, offering support that these are both novel chemotypes active against malaria. Though the piperazine ring was fairly common in both datasets, this shouldn't be of too much concern given that it is a small heterocycle, and quite often was only a small subunit of much larger compounds. Of particular relevance is the presence of the benzothiazole chemotype in both datasets. As is discussed during the literature review to follow (section 4.4), the benzothiazole chemotype has been shown to have widespread applications, including in the field of antimalarial drug design, therefore its presence in these datasets was not surprising. Though this may call into question the novelty of the chemotype with regard to its

antimalarial potential, there has thus far been no mention of it acting as a selective bc_1 inhibitor, therefore the chemotype is still highly promising. Furthermore, the compounds which contained benzothiazole all had alternative side chains and substitution patterns to those present in VS09.

A fingerprint similarity search of the five active compounds was performed against the GSK and St Jude's databases, using the *'Fingerprint Similarity Search Protocol'* as described in the Experimental Chapter. The FCFP_4[36, 37] molecular fingerprint method was used to perform the similarity search, with similarity assessed using the Tanimoto coefficient.[38] Only compounds with a similarity value equal to or greater than 0.5 were reported. This cut off was chosen as it offered a midpoint between similarity and dissimilarity. Also, it was lower than the cut off of 0.7 which was used during LBVS, and given that this was used to represent sufficient similarity, anything lower might suggest the opposite. The number of hits which had a Tanimoto coefficient value greater than 0.5 are reported in table 4.8.

Table. 4.8 Frequency of chemotype hits in the GSK and St Jude's databases.

	Frequency	
Active Compound	**GSK (13,519 compounds)**	**St Jude's (1,236 compounds)**
VS01	0	0
VS09	0	0
VS10	0	0
VS16	3	0
VS18	3	1

As can be seen, for VS01, VS09 and VS10, that is the pyrrolopyrimidine-2,4-dione, benzothiazole and isoalloxazine chemotypes there were no hits in either chemical library with a Tanimoto coefficient value greater than or equal to 0.5, greatly supporting the novelty of these active hits. So, although the benzothiazole core was found to be fairly well reported in both libraries, the quantitative similarity of these compounds compared to VS09 appeared to be sufficiently low, making VS09 a

unique and promising compound. There were a few hits for the piperazine containing compounds (VS16 and VS18), but considering their linear structures this is hardly surprising. Also, these two compounds don't have particularly distinct chemotypes like the other actives, making them the least attractive of the hits for future optimisation.

Once the novelty of the 5 hits had been considered, the active compounds were each investigated in turn. Their activity profiles will now be discussed together with any precedent they showed across the literature, as well as their possible modes of action. Their potential for future investigation will also be discussed.

4.3 VS01

VS01 (fig. 4.4) consists of a pyrrolopyrimidine-2,4-dione core with phenyl substituent's. It was shown to have promising 3D7 inhibition (IC_{50} of 5.24 μM) with little or no NDH2 and bovine bc_1 activity, as well as a reasonable BEI value of 15.3. Though the synthesis of pyrrolopyrimidine-2,4-dione compounds has been of some interest,[39, 40] their only reported pharmacological use has been as agonists and antagonists of adenosine receptors that belong to the superfamily of GPCRs, in particular the adenosine A_{2B} receptor.[41] It is thought that A_{2B} antagonists may have potential clinical applications, as this receptor is implicated in both inflammation and asthma.[42, 43]

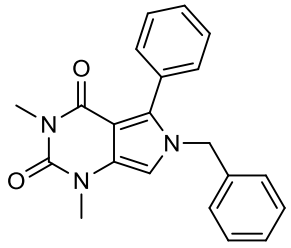
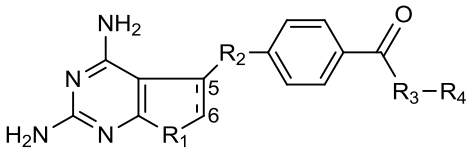


Fig. 4.4 VS01.

A literature search showed there to be no previous reporting of pyrrolopyrimidine-2,4-diones being used as antimalarials, suggesting that this is indeed a novel chemotype suitable for antimalarial investigation. There was however a single publication which reported the use of pyrrolopyrimidines of the general formula shown in figure 4.5, as potentially useful for the treatment of highly drug resistant mutant strains of *Pf*DHFR.[44] To refresh, antifolate compounds such as the prodrug proguanil (fig. 1.12) have been used for the treatment and prophylaxis of malaria,[45-47] as they inhibit the crucial enzyme DHFR. This is required by *plasmodia* for the synthesis of tetrahydrofolic acid, which forms a vital part of the folate biosynthetic pathway, essential for the production of certain nucleotides and amino acids. However, emerging resistance due to an accumulation of mutations in the *dhfr* gene has led to a need for new DHFR inhibitors.[48, 49] If pyrrolopyrimidine containing compounds are capable of circumventing this resistance, then perhaps the pyrrolopyrimidine-2,4-dione template exemplified by VS01, may afford an alternative avenue of investigation, and perhaps go some way to explain the activity of this hit, should it later be found to be inactive towards *Pf*bc$_1$.



Pyrrolo[2,3-d]pyrimidine antifolate

Fig. 4.5 General pyrrolopyrimidine antifolate structure. (R. K. B. Brobey, M. Iwakura, F. Itoh, K. Aso and T. Horii, Parasitol. Int., 1998, **47**, 69-78.)

A possible area of interest for the pyrrolopyrimidine-2,4-dione chemotype would be to investigate how alterations in the side chain affect its *in vitro* activity. From this a SAR study could be developed to potentially improve its potency. Further biological

study of VS01 and other pyrrolopyrimidine-2,4-diones may also yield further insight into the potential mode of action for these compounds.

4.4 VS09

VS09 (fig. 4.6) contains a benzothiazole connected to an imidazole alcohol group via an amine linker. VS09 reported the best activity of the five compounds, with an IC_{50} value against 3D7 of 4.53 ± 1.86 μM. Additionally, VS09 proved to be the most consistent *in vitro*, and displayed little activity towards NDH2, as well as bovine bc_1, suggesting that *Pf*bc_1 may well be its target site. Its BEI value was also found to be the highest of the five actives, which at 23.0 made it highly comparable to that of CQ (24.5), and therefore a potent compound given its relatively small size.

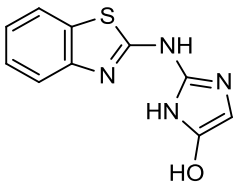


Fig. 4.6 VS09.

The benzothiazole scaffold has been widely reported across the literature with many applications. It can be found in the dye thioflavin which is used in the study of protein aggregation, particularly in Alzheimer's disease patients.[50] It is also found in riluzole,[51] a drug for the treatment of motor neurone disease, as well as in nature in benzothiazole alkaloids.[52] Earlier, the use of virtual screening to identify novel and selective PTR1 inhibitors for the treatment of HAT was discussed, in particular the use of 100 μM to define what constitutes a hit.[13] Two novel chemical series were identified from this study, one of which contained the benzothiazole scaffold, the other a benzimidazole. Further study of these series led to the proposition of their binding modes, as well as low nM PTR1 inhibitors.

The benzothiazole heterocyclic structure clearly has widespread applications, but it has also been previously noted for its potential use in antimalarial drug design. Though there are only a few instances of benzothiazole being used in antimalarials, the first paper reported its possible interest as far back as the late 1960's.[53, 54] In this paper, *Plasmodium berghei* (*P. berghei*),[55] which is a *plasmodium* species that affects African murine rodents and is of no direct concern to man, was used as a surrogate to rapidly test compounds to identify new antimalarials. This practice is of wide interest as it offers a model from which human malaria can be studied, together with analysis of malaria genes using genetic modification.[56] A series of synthesised benzothiazole's were tested for activity against *P. berghei* mice, with several amino alcohols showing weak antimalarial activity. However, activity was only found to occur at toxic doses.

Later work detailed the study of dioxobenzothiazoles as potential antimetabolites of coenzyme Q.[57] Though one compound (fig. 4.7) was shown to have good prophylactic activity without toxicity against *Plasmodium gallinaccum* in chicks, a form of *plasmodium* that causes malaria in poultry, work moved away from the study of benzothiazole's and towards the investigation of alterative bicycloheterocyclic quinones.[58] This was in the hope of elucidating the nature of inhibition at the enzymatic site of CoQ.

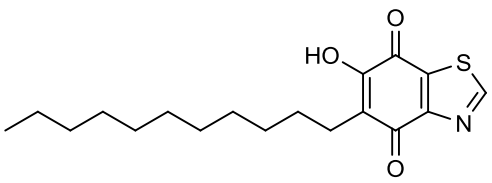


Fig. 4.7 5-n-undecyl-6-hydroxy-4,7-dioxobenzothiazole.

More recently the benzothiazole moiety has been incorporated into much larger structures active against malaria in the single μM range (i.e. fig. 4.8).[59, 60] *De novo*

design was used to develop a new class of non-peptidic inhibitors active against malarial aspartic protease plasmepsin II (PMII). Plasmepsins are a class of enzyme produced by *plasmodia*, and given that their haemoglobin degradation activity is an important cause of symptoms in malaria suffers, plasmepsins therefore pose an interesting target for antimalarial drug design.[61] The benzothiazole moiety was included in the compounds to residue within a hydrophobic pocket of the PMII active site, forming interactions thought to be crucial to its antimalarial activity.

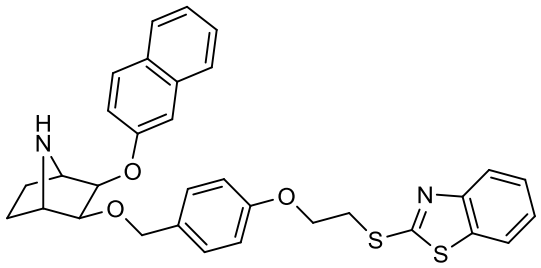


Fig. 4.8 Benzothiazole containing PMII inhibitor.

Perhaps the most important occurrence of the benzothiazole chemotype in the literature was in a series of 2-substituted 6-nitro- and 6-amino-benzothiazoles which were synthesised and tested against *P. falciparum*.[62] The most promising candidates from *in vitro* studies to assess activity and toxicity were carried forward into *in vivo* assays on *P. berghei* infected mice. From this, two had specific antimalarial properties (fig. 4.9) and were active against all stages of the parasite. They were also active on the mitochondrial membrane potential (MMP), causing a drop in MMP that could alter the respiratory chain and lead to parasite death, similar to ATOV. There was however, no discussion about which specific enzymes they may have been inhibiting to produce this drop. Also these compounds were active against multiple pathways (i.e. MMP; haem crystallisation), which may allow for increased efficacy and the reduced development of resistance. Both compounds were deemed worthy of further chemical and biological investigation, though these results have yet to be

published. This work is highly supportive of VS09, and shows the potential which lies within the benzothiazole heterocycle for antimalarial design. Though these compounds do affect MMP, there is no mention as to a particular enzyme they might target, and they are clearly not selective. Thus there are currently no known *Pf*bc$_1$ inhibitors containing the benzothiazole scaffold.

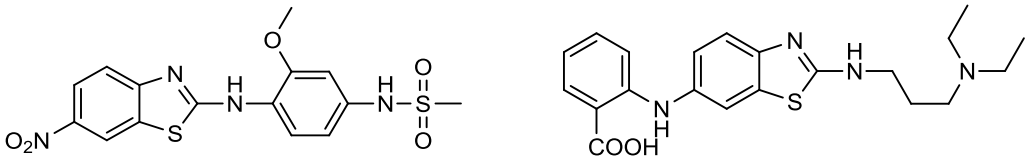


Fig. 4.9 2-substituted 6-nitro- and 6-amino-benzothiazoles.

The benzothiazole template has also appeared in a number of compounds active against the K1 CQR strain of *P. falciparum*, which were later found to inhibit *P. berghei* in mice.[63] However, the most recent appearance of benzothiazole in the literature was concerned with the synthesis of a series of analogues that were designed based on a SBVS study to identify compounds that inhibited falcipain cysteine proteases.[64] The cysteine proteases of *P. falciparum* are collectively known as falcipains,[65] and are responsible for the host haemoglobin hydrolysis. Their inhibition therefore results in parasite cell death, making this a viable antimalarial target. Though the compounds (i.e. fig. 4.10) showed activity against CQR parasites, their IC_{50} values were all fairly poor, with many greater than 50 μM. This would suggests that there is still much room to optimise and refine compounds that contain the benzothiazole chemotype.

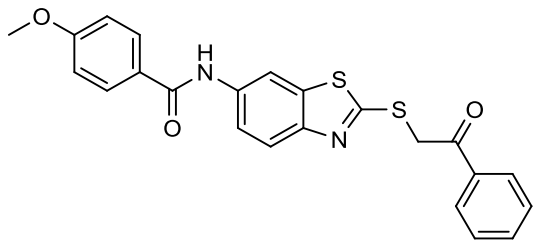


Fig. 4.10 Falcipain inhibitor containing benzothiazole.

Given that the benzothiazole heterocycle has already proven to be of much biological interest, both as an antimalarial structure and in other areas, it looks to be a highly promising lead chemotype to take from the LBVS work performed. By combining the encouraging testing results collected so far, along with earlier discussions in the literature about the possibility that benzothiazole containing compounds may act as antimetabolites of CoQ, the chemotype certainly warrants further investigation to determine its potential as a *Pf*bc$_1$ inhibitor. It's presence amongst the literature with regard to alternative malaria targets is also encouraging, as it may suggest a strong pharmacological profile for VS09. Also, though the benzothiazole chemotype is present in the literature, the papers are all concerned with extensions/side chains unlike that of VS09, whose novelty lies not only in its inception, but also in the combination of the substituted imidazole group, as much as it does in the presence of benzothiazole itself. There is clearly much scope with regard to chemical optimisation of the benzothiazole chemotype, and indeed VS09 itself.

4.5 VS10

VS10 (fig. 4.11) contains a rigid, aromatic tricyclic structure comprised of a pteridine-2,4-dione structure attached to a benzene ring, with a side chain branching out from one of the central amines. This is commonly referred to as an isoalloxazine ring. VS10 was consistently active against 3D7, with an IC_{50} value of 6.41 ± 1.73

μM. It was also poorly active against both NDH2 and bovine bc_1, making it a potentially promising compound, although its BEI was lower than that of CQ at 15.5.

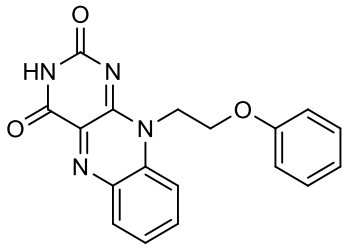


Fig. 4.11 VS10.

The pteridine moiety observed in VS10 is crucial to the folate biosynthetic pathway, and is the core chemotype present in folic acid, which is needed to build and repair DNA. Pteridine has therefore been of much interest with regard to developing potent and selective inhibitors of DHFR.[66, 67] Given that pteridine derivatives compete in the active site of folate enzymes, pteridine containing compounds may also potentiate the activity of other DHFR inhibitors.[68] The chemotype is also present in methotrexate (fig. 4.12), a well known antimetabolite and antifolate drug used in the treatment of cancer and autoimmune diseases.[69] These observations would suggest that VS10 could potentially be a DHFR inhibitor, an avenue which certainly warrants future exploration.

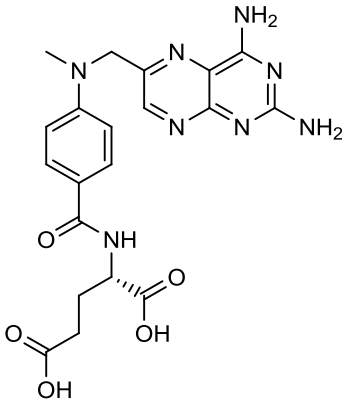


Fig. 4.12 Methotrexate.

The isoalloxazine ring highly resembles that of flavin mononucleotide (FMN, fig. 4.13), which plays a critical role in the ETC. It acts as a prosthetic group for various oxidoreductases, including NADH dehydrogenase, through its reversible inter-conversion of its oxidised (FMN), semiquinone, and reduced ($FMNH_2$) forms.

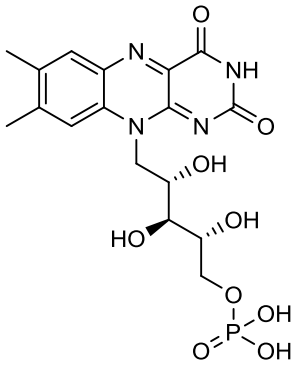


Fig. 4.13 Flavin mononucleotide (FMN).

The ETC is a complicated process and has been discussed in Chapter I.[21, 70-72] It begins with the oxidation of ubiquinone (Q) to ubiquinol (QH2), which is catalysed by NADH:quinone oxidoreductase. NADH binds to complex I, transferring two electrons to FMN, thereby reducing it to $FMNH_2$. The electron accepting isoalloxazine ring of FMN is identical to that of flavin adenine dinucleotide (FAD, fig. 4.14), thus making VS10 structurally related to both FMN, and FAD. FAD is formed at complex II (SDH) through oxidation of $FADH_2$, in a step vital to feed electrons along the mtETC to complex III.[73] $FMNH_2$ is then oxidised in two, one electron steps via a semiquinone intermediate. Each electron transfers from $FMNH_2$ to an Fe-S cluster, and then from the Fe-S cluster to Q, ultimately reducing it to QH2. As a result, four protons translocate across the membrane producing a proton gradient. It has been noted that FMN is a stronger oxidising agent than NAD^+, being particularly useful as it can take part in both one, and two electron transfers.

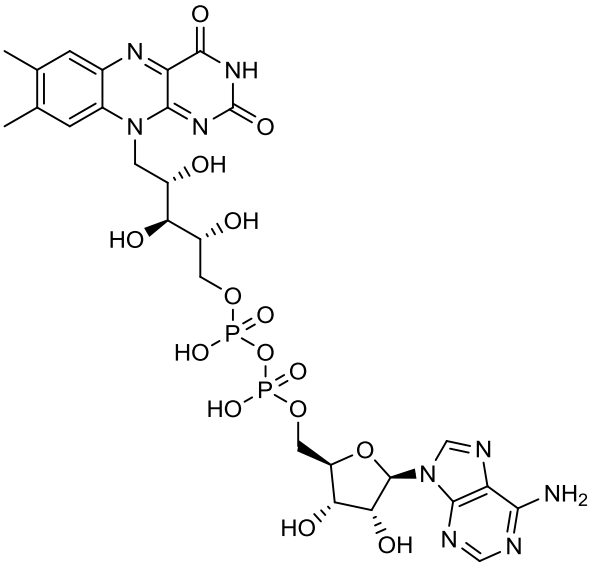


Fig. 4.14 Flavin adenine dinucleotide (FAD).

Upon initial consideration you would expect that a compound containing the isoalloxazine chemotype, whose chemical structure is clearly essential to the ETC (i.e. FMN and FAD), would make for an undesirable template for lead optimisation. The presence of the isoalloxazine ring in VS10 could have even potentially promoted activity of the ETC. However, testing results show this is not the case, being that it had good, reproducible active against 3D7, and observed little toxicity (according to its bovine bc_1 activity). The novelty of VS10 as an antimalarial compound must therefore lie in its side chain, which contains an alkyl chain terminating with a benzene group via an oxygen linker. All things considered, though the aim was to identify chemotypes active against complex III, VS10 may actual elicit its response through competitive inhibition of either one or both of FMN and/or FAD, at complexes I and II respectively. This would still however, block electron transfer, thus inhibiting the ETC. Further biological testing is required to validate this hypothesis. Alteration of the side chain may therefore make for an interesting future study of the isoalloxazine chemotype.

4.6 VS16 and VS18

VS16 and VS18 (fig. 4.15) are perhaps the least interesting chemical structure of the five hits, given that they do not have traditional chemotypes amenable to chemical optimisation. Both contain a piperazine ring with varying functionality branching from either amine. The 3D7 IC_{50} values for these compounds were fairly good, being 5.39 μM and 8.18 μM respectively, yet the results were not reproducible, slightly hampering their potential. Also, though not massive, VS16 observed the highest inhibition of the five active compounds against bovine bc_1 (10%), perhaps further limiting its potential owing to off target toxicity. Both compounds did however have reasonable BEI values of 17.3 and 15.9 respectively, though these were lower than that of CQ, and indeed some of the other hits.

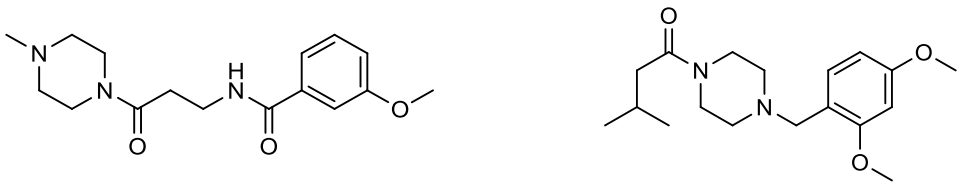


Fig. 4.15 VS16 and VS18.

Piperazine containing antimalarials have been noted amongst the literature, with one paper describing a number of aryl piperazine derivatives active against both CQS and CQR strains of malaria in μM concentrations.[74] More recently, this series was used to develop statistically significant QSARs models, which were later found to successfully predict the activity of unknown compounds.[75] There are also several other mentions of aryl piperazine derivative being active against malaria.[76, 77] However, given the small nature of the piperazine chemotype, it often tends to form part of much larger compounds. An example of this involved the hit to lead optimisation of an imidazolopiperazine containing compound (fig. 4.16) active

against sensitive and resistant parasite strains.[11] The lead compounds from this study ultimately showed good potencies *in vitro*, as well as decent oral exposure levels *in vivo*.

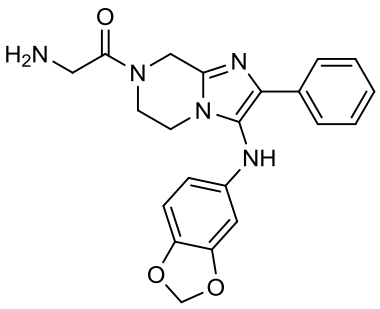


Fig. 4.16 Imidazolopiperazine compound.

Despite their inconsistent performance in the 3D7 bioassay, the literature would suggest that there is still potential within the piperazine heterocycle for use in antimalarial drug design. Perhaps further development and optimisation of the side chains of VS16 and VS18 may yield some improvements in their activity profiles.

4.7 Summary of Testing/Analysis

From the results presented in this chapter it can be concluded that five of the compounds identified by LBVS were active against the 3D7 strain of the malaria parasite, with each containing a novel structural chemotype. Though the aim of LBVS was to identify novel *Pf*bc$_1$ inhibitors suitable for chemical optimisation, further *in vitro* testing is required (when it becomes available) in order to validate this claim. However, the results collected thus far look highly promising, with the benzothiazole chemotype being of particular interest. It would therefore be possible to conclude that LBVS was successful in its aim of using existing chemical information to inform drug development with regard to antimalarial design.

Ultimately, these novel chemotypes will go on to form the basis of another iteration of the molecular design loop, to continually refine the discovery process.

The 5 active hits have also been the subject of a molecular docking study, the results of which are discussed in the following chapter. These results were used to further rationalise their potential as *Pf*bc_1 inhibitors. Chapter V also reports the findings from a number of other docking studies, which were performed to rationalise/explain the observed activity patterns of other compounds active against the bc_1 complex.

4.8 References

1. P. Workman, *Curr. Pharm. Design*, 2003, **9**, 891-902.
2. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
3. N. Fisher, C. K. Castleden, I. Bourges, G. Brasseur, G. Dujardin and B. Meunier, *J. Biol. Chem.*, 2004, **279**, 12951-12958.
4. D. W. Wilson, B. S. Crabb and J. G. Beeson, *Malaria Journal*, 2010, **9**.
5. N. Fisher, A. J. Warman, S. A. Ward and G. A. Biagini, in *Methods in Enzymology, Vol 456*, ed. W. S. Allison, Elsevier Academic Press Inc, San Diego, Editon edn., 2009, vol. 456, pp. 303-320.
6. G. L. Patrick, *An Introduction to Medicinal Chemistry*, Oxford University Press, 2005.
7. R. P. Hertzberg and A. J. Pope, *Curr. Opin. Chem. Biol.*, 2000, **4**, 445-451.
8. J. N. Burrows, K. Chibale and T. N. C. Wells, *Curr. Top. Med. Chem.*, 2011, **11**, 1226-1254.
9. P. B. McKay, M. B. Peters, G. Carta, C. T. Flood, E. Dempsey, A. Bell, C. Berry, D. G. Lloyd and D. Fayne, *Bioorganic and Medicinal Chemistry Letters*, 2011, **21**, 3335-3341.
10. T. Wu, A. Nagle, T. Sakata, K. Henson, R. Borboa, Z. Chen, K. Kuhen, D. Plouffe, E. Winzeler, F. Adrian, T. Tuntland, J. Chang, S. Simerson, S. Howard, J. Ek, J. Isbell, X. Deng, N. S. Gray, D. C. Tully and A. K. Chatterjee, *Bioorganic and Medicinal Chemistry Letters*, 2009, **19**, 6970-6974.
11. T. Wu, A. Nagle, K. Kuhen, K. Gagaring, R. Borboa, C. Francek, Z. Chen, D. Plouffe, A. Goh, S. B. Lakshminarayana, J. Wu, H. Q. Ang, P. Zeng, M. L. Kang, W. Tan, M. Tan, N. Ye, X. Lin, C. Caldwell, J. Ek, S. Skolnik, F. Liu, J. Wang, J. Chang, C. Li, T. Hollenbeck, T. Tuntland, J. Isbell, C. Fischli, R. Brun, M. Rottmann, V. Dartois, T. Keller, T. Diagana, E. Winzeler, R. Glynne, D. C. Tully and A. K. Chatterjee, *Journal of Medicinal Chemistry*, 2011, **54**, 5116-5130.
12. F. P. Da Cruz, C. Martin, K. Buchholz, M. J. Lafuente-Monasterio, T. Rodrigues, B. Sönnichsen, R. Moreira, F. J. Gamo, M. Marti, M. M. Mota, M. Hannus and M. Prudêncio, *J. Infect. Dis.*, 2012, **205**, 1278-1286.
13. C. P. Mpamhanga, D. Spinks, L. B. Tulloch, E. J. Shanks, D. A. Robinson, I. T. Collie, A. H. Fairlamb, P. G. Wyatt, J. A. Frearson, W. N. Hunter, I. H. Gilbert and R. Brenk, *Journal of Medicinal Chemistry*, 2009, **52**, 4454-4465.
14. K. J. Simmons, I. Chopra and C. W. G. Fishwick, *Nat. Rev. Microbiol.*, 2010, **8**, 501-510.
15. A. Dove, *Nat. Biotechnol.*, 1999, **17**, 859-863.
16. J. Hüser, *High-Throughput Screening in Drug Discovery*, Wiley-VCH, 2006.
17. A. Bender and R. C. Glen, *Journal of Chemical Information and Modeling*, 2005, **45**, 1369-1375.
18. C. K. Dong, V. Patel, J. C. Yang, J. D. Dvorin, M. T. Duraisingh, J. Clardy and D. F. Wirth, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 972-975.
19. S. C. Leung, P. Gibbons, R. Amewu, G. L. Nixon, C. Pidathala, W. D. Hong, B. Pacorel, N. G. Berry, R. Sharma, P. A. Stocks, A. Srivastava, A. E. Shone, S. Charoensutthivarakul, L. Taylor, O. Berger, A. Mbekeani, A. Hill, N. E. Fisher, A. J. Warman, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Journal of Medicinal Chemistry*, 2012, **55**, 1844-1857.
20. C. Pidathala, R. Amewu, B. Pacorel, G. L. Nixon, P. Gibbons, W. D. Hong, S. C. Leung, N. G. Berry, R. Sharma, P. A. Stocks, A. Srivastava, A. E. Shone, S. Charoensutthivarakul, L. Taylor, O. Berger, A. Mbekeani, A. Hill, N. E. Fisher, A. J. Warman, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Journal of Medicinal Chemistry*, 2012, **55**, 1831-1843.
21. A. R. Crofts, *Annu. Rev. Physiol.*, 2004, **66**, 689-733.
22. D. M. Kramer, A. G. Roberts, F. Muller, J. Cape and M. K. Bowman, in *Methods in Enzymology*, eds. S. Helmut and P. Lester, Academic Press, Editon edn., 2004, vol. Volume 382, pp. 21-45.
23. R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. Oneill, *MedChemComm*, 2012, **3**.
24. S. E. Hampton, B. Baragaña, A. Schipani, C. Bosch-Navarrete, J. A. Musso-Buendía, E. Recio, M. Kaiser, J. L. Whittingham, S. M. Roberts, M. Shevtsov, J. A. Brannigan, P. Kahnberg, R. Brun, K. S. Wilson, D. González-Pacanowska, N. G. Johansson and I. H. Gilbert, *ChemMedChem*, 2011, **6**, 1816-1831.
25. L. R. Whittell, K. T. Batty, R. P. M. Wong, E. M. Bolitho, S. A. Fox, T. M. E. Davis and P. E. Murray, *Bioorganic and Medicinal Chemistry*, 2011, **19**, 7519-7525.

26. Y. Zhang, M. Anderson, J. L. Weisman, M. Lu, C. J. Choy, V. A. Boyd, J. Price, M. Sigal, J. Clark, M. Connelly, F. Zhu, W. A. Guiguemde, C. Jeffries, L. Yang, A. Lemoff, A. P. Liou, T. R. Webb, J. L. Derisi and R. K. Guy, *ACS Medicinal Chemistry Letters*, 2010, **1**, 460-465.
27. I. D. Kuntz, K. Chen, K. A. Sharp and P. A. Kollman, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 9997-10002.
28. C. Abad-Zapatero and J. T. Metz, *Drug Discovery Today*, 2005, **10**, 464-469.
29. A. L. Hopkins, C. R. Groom and A. Alex, *Drug Discovery Today*, 2004, **9**, 430-431.
30. C. Abad-Zapatero, *Expert. Opin. Drug Discov.*, 2007, **2**, 469-488.
31. N. C. P. Araújo, V. Barton, M. Jones, P. A. Stocks, S. A. Ward, J. Davies, P. G. Bray, A. E. Shone, M. L. S. Cristiano and P. M. O'Neill, *Bioorganic & Medicinal Chemistry Letters*, 2009, **19**, 2038-2043.
32. M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis and P. D. Leeson, *Journal of Medicinal Chemistry*, 2003, **46**, 1250-1256.
33. M. Vieth, M. G. Siegel, R. E. Higgs, I. A. Watson, D. H. Robertson, K. A. Savin, G. L. Durst and P. A. Hipskind, *Journal of Medicinal Chemistry*, 2004, **47**, 224-232.
34. F. J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J. L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon and J. F. Garcia-Bustos, *Nature*, 2010, **465**, 305-U356.
35. W. A. Guiguemde, A. A. Shelat, D. Bouck, S. Duffy, G. J. Crowther, P. H. Davis, D. C. Smithson, M. Connelly, J. Clark, F. Y. Zhu, M. B. Jimenez-Diaz, M. S. Martinez, E. B. Wilson, A. K. Tripathi, J. Gut, E. R. Sharlow, I. Bathurst, F. El Mazouni, J. W. Fowble, I. Forquer, P. L. McGinley, S. Castro, I. Angulo-Barturen, S. Ferrer, P. J. Rosenthal, J. L. DeRisi, D. J. Sullivan, J. S. Lazo, D. S. Roos, M. K. Riscoe, M. A. Phillips, P. K. Rathod, W. C. Van Voorhis, V. M. Avery and R. K. Guy, *Nature*, 2010, **465**, 311-315.
36. D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742-754.
37. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *Journal of Chemical Information and Computer Sciences*, 2002, **42**, 1273-1280.
38. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
39. K. Hirota, Y. Yamada, T. Asao and S. Senda, *Chemical and Pharmaceutical Bulletin*, 1981, **29**, 1525-1532.
40. N. M. Sekhar, P. V. R. Acharyulu and Y. Anjaneyulu, *Tetrahedron Lett.*, 2011, **52**, 4140-4144.
41. M. W. Beukers, I. Meurs and A. P. Ijzerman, *Med. Res. Rev.*, 2006, **26**, 667-698.
42. W. A. Sands and T. M. Palmer, *Immunol. Lett.*, 2005, **101**, 1-11.
43. M. Livingston, L. G. Heaney and M. Ennis, *Inflamm. Res.*, 2004, **53**, 171-178.
44. R. K. B. Brobey, M. Iwakura, F. Itoh, K. Aso and T. Horii, *Parasitol. Int.*, 1998, **47**, 69-78.
45. M. Schlitzer, *ChemMedChem*, 2007, **2**, 944-986.
46. D. C. M. Chan and A. C. Anderson, *Curr. Med. Chem.*, 2006, **13**, 377-398.
47. A. Nzila, *J. Antimicrob. Chemother.*, 2006, **57**, 1043-1054.
48. J. Yuvaniyama, P. Chitnumsub, S. Kamchonwongpaisan, J. Vanichtanankul, W. Sirawaraporn, P. Taylor, M. D. Walkinshaw and Y. Yuthavong, *Nat. Struct. Biol.*, 2003, **10**, 357-365.
49. P. K. Rathod and M. A. Phillips, *Nat. Struct. Biol.*, 2003, **10**, 316-318.
50. L. S. Wolfe, M. F. Calabrese, A. Nath, D. V. Blaho, A. D. Miranker and Y. Xiong, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 16863-16868.
51. J. Wokke, *The Lancet*, 1996, **348**, 795-799.
52. L. Le Bozec and C. J. Moody, *Australian Journal of Chemistry*, 2009, **62**, 639-647.
53. A. Burger and S. N. Sawhney, *Journal of Medicinal Chemistry*, 1968, **11**, 270-&.
54. D. M. Aviado, *Exp. Parasitol.*, 1969, **25**, 399-482.
55. I. H. Vincke and M. Lips, *Annales de la Societe belge de medecine tropicale*, 1948, **28**, 97-104.
56. C. J. Janse, J. Ramesar and A. P. Waters, *Nat. Protoc.*, 2006, **1**, 346-356.
57. M. D. Friedman, P. L. Stotter, T. H. Porter and K. Folkers, *Journal of Medicinal Chemistry*, 1973, **16**, 1314-1316.
58. T. H. Porter and K. Folkers, *ANTIMETABOLITEN DES COENZYMS Q. MOGLICHKEITEN IHRER ANWENDUNG ALS ANTIMALARIA MITTEL*, 1974, **86**, 635-645.
59. D. A. Carcache, S. R. Hörtner, A. Bertogg, F. Diederich, A. Dorn, H. P. Märki, C. Binkert and D. Bur, *Helvetica Chimica Acta*, 2003, **86**, 2192-2209.
60. D. A. Carcache, S. R. Hortner, P. Seiler, F. Diederich, A. Dorn, H. P. Marki, C. Binkert and D. Bur, *Helvetica Chimica Acta*, 2003, **86**, 2173-2191.

61. A. M. Silva, A. Y. Lee, S. V. Gulnik, P. Majer, J. Collins, T. N. Bhat, P. J. Collins, R. E. Cachau, K. E. Luker, I. Y. Gluzman, S. E. Francis, A. Oksman, D. E. Goldberg and J. W. Erickson, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 10034-10039.
62. S. Hout, N. Azas, A. Darque, M. Robin, C. Di Giorgio, M. Gasquet, J. Galy and P. Timon-David, *Parasitology*, 2004, **129**, 525-535.
63. K. Pudhom, K. Kasai, H. Terauchi, H. Inoue, M. Kaiser, R. Brun, M. Ihara and K. Takasu, *Bioorganic & Medicinal Chemistry*, 2006, **14**, 8550-8563.
64. F. Shah, Y. Wu, J. Gut, Y. Pedduri, J. Legac, P. J. Rosenthal and M. A. Avery, *MedChemComm*, 2011, **2**, 1201-1207.
65. P. J. Rosenthal, *Int. J. Parasit.*, 2004, **34**, 1489-1499.
66. A. C. Anderson, *Drug Discovery Today*, 2005, **10**, 121-128.
67. I. H. Gilbert, *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 2002, **1587**, 249-257.
68. A. Nzila, *Drug Discovery Today*, 2006, **11**, 939-944.
69. F. M. Huennekens, *Advances in Enzyme Regulation*, 1994, **34**, 397-419.
70. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
71. C. C. Wang, *Journal of Medicinal Chemistry*, 1984, **27**, 1-9.
72. P. L. Olliaro and Y. Yuthavong, *Pharmacol. Ther.*, 1999, **81**, 91-110.
73. N. Suraveratum, S. R. Krungkrai, P. Leangaramgul, P. Prapunwattana and J. Krungkrai, *Mol. Biochem. Parasitol.*, 2000, **105**, 215-222.
74. C. A. Molyneaux, M. Krugliak, H. Ginsburg and K. Chibale, *Biochem. Pharmacol.*, 2005, **71**, 61-68.
75. E. Ibezim, P. R. Duchowicz, E. V. Ortiz and E. A. Castro, *Chemometrics Intell. Lab. Syst.*, 2012, **110**, 81-88.
76. A. Mendoza, S. Pérez-Silanes, M. Quiliano, A. Pabón, S. Galiano, G. González, G. Garavito, M. Zimic, A. Vaisberg, I. Aldana, A. Monge and E. Deharo, *Exp. Parasitol.*, 2011, **128**, 97-103.
77. C.-A. Molyneaux, M. Krugliak, H. Ginsburg and K. Chibale, *Biochem. Pharmacol.*, 2005, **71**, 61-68.

Chapter V

Molecular Docking Studies

5. Molecular Docking Studies 231

5.1 Scoring Functions 232

5.1.1 GOLDScore 234

5.1.2 ChemScore 235

5.2 Cytochrome bc_1 Protein Target 237

5.2.1 The Q_o Site 239

5.2.1.1 Stigmatellin 240

5.2.1.2 Q_o Binding Mode 242

5.2.2 Developing a Docking Protocol 246

5.2.2.1 Docking of Stigmatellin 247

5.2.2.2 Docking of Atovaquone 253

5.2.2.3 Docking of Quinolone Esters 260

5.2.2.4 Docking of Lead Quinolone Compound 269

5.2.3 The Q_i Site 272

5.2.3.1 Docking of Antimycin A 272

5.2.3.2 Docking of NQNO 276

5.2.3.3 Docking of HDQ 281

5.2.4 Docking of LBVS Hits 286

5.3 Summary of Docking Studies 293

5.4 References 295

5. Molecular Docking Studies

Predicting the binding modes and affinities of compounds when they interact with a protein binding site lies at the heart of structure based drug design.[1] Molecular docking and its uses have already been reviewed in Chapter I, but in essence, its aim is to predict the preferred binding orientation of one molecule with respect to a second.[2] Protein-ligand docking is of massive importance in rational drug design, as it enables the prediction of the binding orientation of small molecules within a protein, which may be used to predict it's binding affinity.[3] A large number of methods are currently available for use in protein-ligand docking,[4-7] including DOCK,[8] FlexX,[9] PRO_LEADS[10] and GOLD,[11, 12] with most approaches considering the protein to be (almost) rigid, and the ligand to be flexible.[6]

The key characteristic of a good docking protocol however, is its ability to reproduce the experimental binding mode of a ligand.[1] To test this, a ligand should be removed from the crystal structure of its protein-ligand complex, and then docked back into its binding site. The docked binding mode is then compared with the experimental binding mode, with RMSD providing a means of measuring the similarity between the actual and predicted poses. The prediction of a binding mode is considered successful if the RMSD is below a certain threshold, usually 2 Å.[1]

The protein-ligand docking described in this thesis was all performed using the docking program GOLD 5.0.1[13] (Genetic Optimisation for Ligand Docking), which searched for the best ligand interaction poses using a genetic algorithm (Chapter I). Work has previously been carried in order to establish the success rate of GOLD in predicting binding poses, using a large and carefully constructed set of protein-ligand

complexes.[14] GOLD was shown to reproduce the native binding pose of a ligand with a 68% success rate, when using the CCDC/Astex validation set of 305 complexes, all of which were taken from the PDB library.[15]

Molecular docking essentially consists of two problems, the first being the necessity of a mechanism for exploring the extremely large "search space" of possible protein-ligand geometries, and the second the need to score and/or rank poses in order to identify the most likely binding modes. Both of these concepts have already been explored during Chapter I, so particular scoring methods will now be introduced.

A docking protocol developed in GOLD[13] consists of three main parts.[11, 12] Firstly, a scoring function must be used in order to rank the different binding modes that are reported. Next, a mechanism is required in order to place the ligand into the binding site. GOLD uses a unique method to do this by adding fitting points to hydrogen bonding groups on the protein and ligand, mapping acceptor points on the ligand to donor points on the protein, and vice versa. GOLD also generates hydrophobic fitting points in the protein cavity, onto which ligand CH groups are mapped. Finally, a search algorithm is employed to explore possible binding modes. GOLD uses a GA in which several parameters are optimised, such as the dihedrals of the ligands rotatable bonds, the ligands ring geometries, dihedrals of protein OH and NH_3^+ groups, as well as the mapping of fitting points.

5.1 Scoring Functions

As it is not always clear which docking protocol will give the best result for a particular problem,[16, 17] it is important to carefully consider each case individually. When several docking poses are generated it is crucial to score or rank these poses using some function related to the free energy of association of the intermolecular

complex. There are a wide range of scoring functions available,[18] and the ability to accurately predict the potency of ligand binding within a protein is of significant value, providing useful starting points for drug discovery.[19, 20] Once the ligands are docked, the resulting interactions can be scored giving a quantitative measure of fit quality. Scoring functions are approximate mathematical methods used to predict the strength of the non-covalent interactions between two molecules after they have been docked (also referred to as binding affinity), and it is common practice to use scoring functions in protein-ligand docking.[21] Scoring functions can be grouped into three categories: force field based, empirical, and knowledge based.[22]

Knowledge based methods rely on the idea that a sufficiently large data sample can serve to derive rules and general principles inherently stored in this knowledge database.[22-26] It is based on statistical observations of intermolecular close contacts in large 3D datasets, such that the interaction potential between each ligand-protein atom pair is calculated as a potential of mean force. The method is founded on the assumption that close intermolecular interactions between certain types of atoms or functional groups that occur more frequently than one would expect by a random distribution, are likely to be energetically favourable and therefore contribute favourably to binding affinity.[27] Knowledge based methods were not employed in this research and will not be discussed further, as these methods could not be used to perform rescoring when a water molecule was present, which was essential for this research. Examples of force field based and empirical methods will now be introduced more fully in the context of GOLD docking.

5.1.1 GOLDScore

GOLDScore[11, 12] is a force field based scoring function that provides a quantitative means of estimating the scoring affinity of a pose, by summing the strength of intermolecular van der Waals and electrostatic interactions between all atoms of the two molecules in the complex. The intramolecular energies of the two binding partners is also frequently considered, and since binding normally takes place in the presence of water, the desolation energies of the ligand and the protein are sometimes taken into account using implicit solvation. Force field based scoring functions are primarily derived from force fields such as AMBER,[28] which are frequently used in molecular dynamics simulations.

The GOLDScore of a pose can be calculated using equation 5.1. The GOLDScore fitness function is comprised of four components: protein-ligand hydrogen bond energy (external H-bond; S_{hb_ext}); protein-ligand van der Waals energy (external vdw; S_{vdw_ext}); intramolecular strain in the ligand vdw energy (internal vdw; S_{vdw_int}); ligand torsional strain energy (internal torsion; S_{tor_int}). A fifth component may also be added, which looks at the ligand intramolecular hydrogen bond energy (internal H-bond), but this is not included by default and its omission has been found to give better results.[1]

$$GOLDScore = S_{hb\_ext} + S_{vdw\_ext} + S_{vdw\_int} + S_{tor\_int}$$

Eq. 5.1 GOLDScore fitness function calculation.

The fitness score is taken as the negative of the sum of the component energy terms so that a larger fitness score suggests a better binding complex. The external vdw term is multiplied by a factor of 1.375 when the total fitness score is computed, and is an empirical correction to encourage protein-ligand hydrophobic contact. The

GOLDScore fitness function has been optimised for the prediction of ligand binding positions rather than for the prediction of binding affinities.

5.1.2 ChemScore

ChemScore is an empirical based scoring method,[10, 29, 30] derived to reproduce experimentally determined complex structures from physicochemical properties, based on counts of the number of various types of interactions between two binding partners.[31] Counting may be based on the number of ligand and receptor atoms in contact with each other, or by calculating the change in solvent accessible surface area in the complex compared to the uncomplexed ligand and protein. The coefficients of the scoring function are usually fitted using MLR methods, and may include contributions from hydrogen bonding, ionic interactions, lipophilic interactions and the loss of internal conformational freedom of the ligand. ChemScore was derived empirically from a set of 82 protein-ligand complexes for which measured binding affinities were available. These were then trained by linear regression against the measured affinity data. The ChemScore function estimates the total free energy of binding change ($\Delta G_{binding}$) that occurs on ligand binding, and is calculated using equation 5.2. In equation 5.2, S_{hbond}, S_{metal} and S_{lipo} represent the scores for hydrogen-bonding, acceptor-metal and lipophilic interactions respectively, whilst H_{rot} provides a measure for the loss of conformational entropy in the ligand upon binding to the protein. The ΔG terms in the equation are coefficients that were derived from the MLR analysis of the 82 protein-ligand complexes (table 5.1).

$$\Delta G_{binding} = \Delta G_o + \Delta G_{hbond} S_{hbond} + \Delta G_{metal} S_{metal} + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} H_{rot}$$

Eq. 5.2 ChemScore free energy of binding calculation.

Table. 5.1 ChemScore free energy of binding parameters.

ΔG_o	-5.480 (kJ/mol)
ΔG_{hbond}	-3.340 (kJ/mol)
ΔG_{metal}	-6.030 (kJ/mol)
ΔG_{lipo}	-0.117 (kJ/mol)
ΔG_{rot}	2.560 (kJ/mol)

The ChemScore function was adapted for docking by adding in a protein-ligand clash penalty term (E_{clash}) and an internal ligand energy torsion term (E_{int}). Together these militate against close contacts in docking and poor internal conformations. Additionally a covalent energy term (E_{cov}) was included, with ChemScore ultimately calculated using equation 5.3.

$$ChemScore = \Delta G_{binding} + E_{clash} + E_{int} + E_{cov}$$

Eq. 5.3 Calculation of ChemScore.

As with GOLDScore, the higher the ChemScore value the better the predicted binding pose. However, GOLDScore and ChemScore are unrelated and as such, are not comparable to one another. Though GOLDScore has been found to perform marginally better at predicting native binding poses for certain complexes,[1] it is difficult to generalise these observations, and thus it is necessary to evaluate the relative merits of different methods for individual examples.

When protein-ligand docking is used to dock large libraries of compounds into a target binding site, the docking of each compound needs to be quick, otherwise fewer binding modes are sampled, potentially reducing the success rates.[1] The ChemScore docking function is up to three times faster than that of GOLDScore, with minimal sacrifice of accuracy.[1]

5.2 Cytochrome bc_1 Protein Target

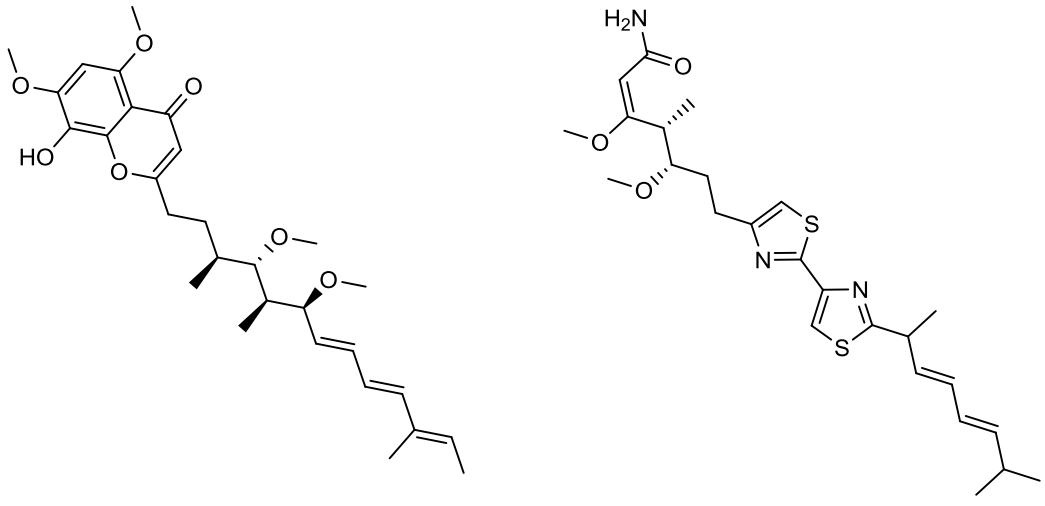
Pfbc_1 has been validated as a target for novel antimalarial drug development through the study of drugs such as ATOV, which is a competitive inhibitor of CoQ.[32, 33] The bc_1 complex is a homodimeric transmembrane protein with a molecular mass of 240 kDa. It contains an electron-transferring core that is comprised of three catalytic subunits (cytochrome b, cytochrome c_1, ISP) that catalyse the transfer of electrons from ubiquinol to cytochrome c, coupled with the translocation of protons across the inner mitochondrial membrane (fig. 5.1).[34, 35] Loss of bc_1 activity results in the loss of mitochondrial function, and the collapse of the transmembrane electrochemical potential.[36, 37]

Fig. 1.16 **(a)** Homodimeric structure of the yeast cytochrome bc_1 complex (PDB accession code 3CX5). **(b)** The structure and Q-cycle mechanism of the catalytic core of the bc_1 complex. (V. Barton, N. Fisher, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Curr. Opin. Chem. Biol.*, 2010, **14**, 440-446.)

The generation of the membrane potential is achieved through a bifurcated redox pathway (Q-cycle) involving the endogenous compounds ubiquinol and ubiquinone, the net result of which is the release of two protons to the cytosolic side of the inner membrane per two electrons transferred from ubiquinol to cytochrome c.[32, 35, 38]

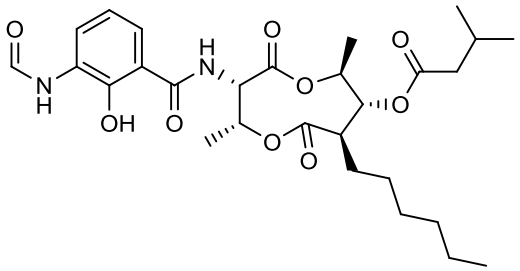
Ubiquinol is oxidised at the Q_o site of the bc_1 complex, which is located at the interface of two four-helical bundles in the cytochrome b subunit, residing in between the b_L heme and the 2Fe-2S cluster of ISP. The Q_o site is on the other side of the membrane to Q_i, linked via a transmembrane electron transfer pathway provided by the membrane spanning cytochrome b subunit and the b_H heme. The Q_i site is responsible for the reduction of ubiquinone to ubiquinol.

bc_1 Inhibitors are traditionally divided into two classes based on their specific sites of action: class I for Q_o inhibitors and class II for Q_i inhibitors.[39, 40] Class I inhibitors such as SMA and myxothiazol (fig. 5.2) inhibit electron bifurcation at the Q_o site, whilst class II inhibitors such as the natural fungicide antimycin A (fig. 5.3), block the electron transfer path from heme b_H to quinone. The use of these compounds is however limited as they are often highly toxic in mammals and other non-pathogenic organisms.



SMA Myxothiazol

Fig. 5.2 Known Q_o inhibitors stigmatellin and myxothiazol.



Antimycin A

Fig. 5.3 Known Q_i inhibitor antimycin A.

Molecular docking has been employed in an attempt to improve our understanding as to how certain compounds are thought to elicit their response against the cytochrome bc_1 complex. An atomic structure for *Pf*bc_1 is not currently available, though work is ongoing to generate a protein structure using homology modelling. *In silico* work was therefore performed largely using the crystal structure of the yeast cytochrome bc_1 protein (PDB accession code 3CX5).[41] Though not identical, the yeast protein shares 40% sequence homology with the parasite bc_1 complex, and is highly conserved across the Q_o region.[42]

5.2.1 The Q_o Site

The Q_o pocket of the bc_1 complex is approximately 15 Å in length and has the shape of a saddle.[39] It is highly hydrophobic and has a relatively wide opening at one end that becomes increasingly narrower moving inwards. The pocket opens up again after the narrow, flat constriction but is sealed at the other end. A number of aromatic and aliphatic residues line the inside of the Q_o pocket and form several key interactions with inhibitors. Q_o inhibitors can be categorised as those which either promote movement, or fix the position of the water soluble domain of ISP.[43] Several

compounds were docked into the Q_o site, thus its structure will be discussed more fully.

5.2.1.1 Stigmatellin

SMA (fig. 5.2) is a natural antifungal agent produced by myxobacteria *S. Aurantiaca*. It is a chromone derivative which fixes the position of the water soluble head group of ISP. In fact, its binding to the Q_o pocket is critically dependent on the presence of the ISP.[44] The water soluble head group of ISP is a histidine amino acid (His181). The protonated N atom of His181 forms hydrogen bonds with the methoxy oxygen and the carbonyl oxygen of the chromone head group of SMA, with distances of 3.5 Å and 3.4 Å respectively (heavy atom to heavy atom distances).[39] There is also a strong hydrogen bond (2.6 Å) between the hydroxyl group of SMA, and an oxygen atom of the glutamic acid (Glu272) amino acid in the Q_o pocket, mediated by a water molecule to form a key hydrogen bonding network. This water molecule is thought to bridge the gap between Glu272 and SMA, stabilising the chromone head group. These two interactions are thought to be the most crucial in explaining the antimalarial activity of Q_o inhibitors, and are illustrated in figure 5.4. Figure 5.5 illustrates the native binding pose of SMA in the yeast crystal structure of cytochrome bc_1. The interaction between SMA and His181 is particularly important as this prevents the oxidation of ubiquinol, and thus inhibits electron transfer.[45]

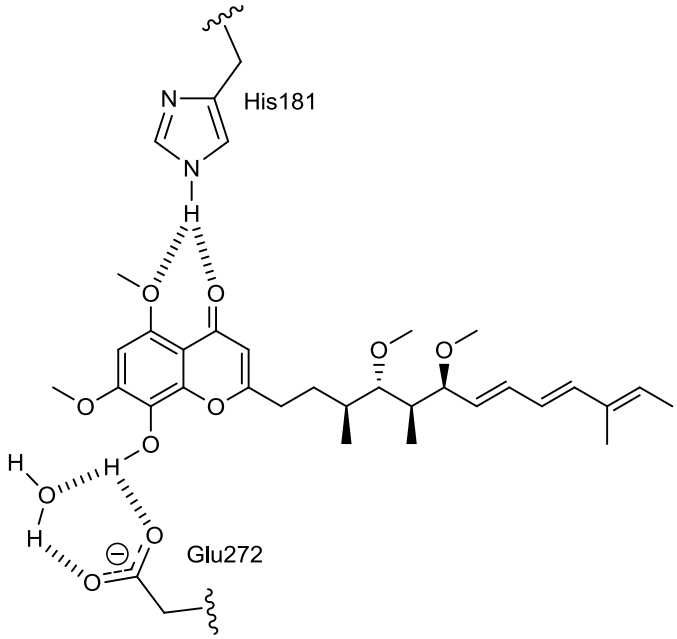


Fig. 5.4 Diagram of the key interactions between SMA in the Q_o pocket, and the His181 and Glu272 amino acids in the cytochrome bc_1 complex.

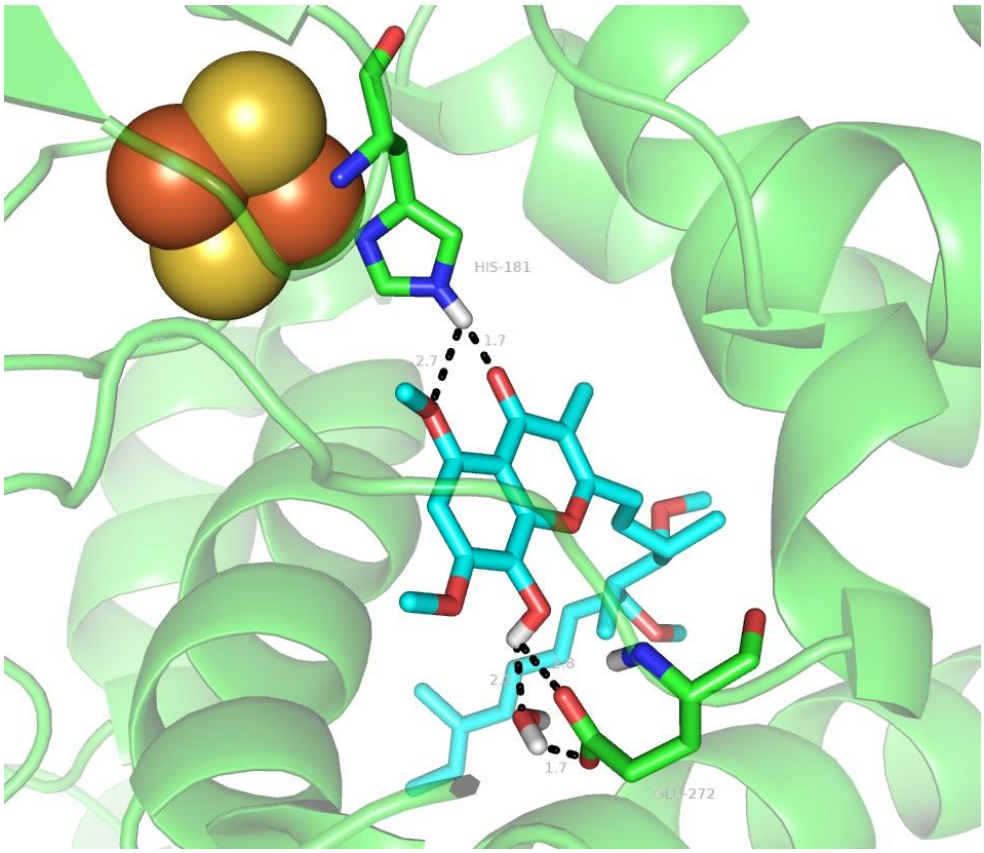


Fig. 5.5 The native binding pose of SMA in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). Two key interactions are observed; one between His181 and the carbonyl and methoxy oxygen atoms of the chromone ring, the other a water mediated H-bonding network between Glu272 and the hydroxyl group. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

There are however a number of additional van der Waals and hydrophobic interactions between the amino acids of the Q_o pocket and SMA, which help to stabilise the compound in the active site. The dimethoxyphenol moiety of the chromone ring of SMA intercalates between proline (Pro270) and isoleucine (Ile146) amino acids, whilst the largely hydrophobic tail interacts with phenylalanine (Phe274), methionine (Met124, Met129), alanine (Ala125) and leucine (Leu294) residues.

As stated above there is good sequence homology between the parasite bc_1 complex and the yeast protein, particularly in the Q_o region.[42] In fact, all of the residues involved in the binding of the head group in this region are fully conserved across all mitochondrial bc_1 complexes.[45, 46] For this reason the yeast protein was perfectly suited for the docking needs of the research undertaken.

5.2.1.2 Q_o Binding Mode

Through structural analysis of the binding of hydroxyquinone inhibitor HHDBT (fig. 5.6) and SMA at the Q_o site, the electron and proton transfer procedures that occur at this site have been deduced.[45]

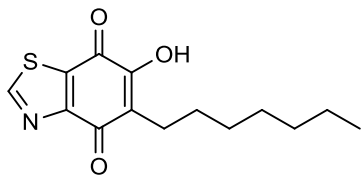


Fig. 5.6 HHDBT Q_o inhibitor.

Figure 5.7 depicts this process. The natural substrate ubiquinol is bound in the Q_o pocket and stabilised with its functional groups pointing towards the primary acceptors of the low and high potential electron transfer chains, heme b_L and the ISP

cluster respectively. Glu272 rotates into the binding pocket and forms a hydrogen bond to the hydroxyl group facing the heme b_L, as visible when SMA is bound at the Q_o site.[47] Since the extrinsic catalytic Rieske domain is mobile and His181 has a pK_a of 7.5 when the cluster is oxidised,[48] a considerable fraction of the latter is not protonated under physiological conditions. Thus, the Rieske domain can be stabilised in the b-position by forming a hydrogen bond between deprotonated His181 and the hydroxyl group of ubiquinol. This brings the ISP cluster into a suitable distance for electron transfer. The initial enzyme-substrate complex is stabilised by hydrogen bonds to the primary ligands, His181 and Glu272, which additionally serve as primary proton acceptors.[47, 49] Ubiquinol is oxidised in a bifurcated manner, transferring two electrons to the ISP cluster and heme b_L. The mechanism and order of events are heavily debated. Some mechanisms assume a sequential reaction in which the first electron reduces the ISP cluster and a relatively unstable semiquinone intermediate reduces heme b_L.[49] A sequential mechanism has also been proposed in which a stable semiquinone is formed and is anti-ferromagnetically coupled to the ISP centre until it is oxidised by heme b_L.[50] The concerted mechanism assumes that neither electron is transferred independently, but rather the semiquinone is so unstable that ubiquinol cannot reduce the ISP centre unless the semiquinone reduces heme b_L.[51, 52] In such a mechanism the concentration of ubisemiquinone is so low as to be almost nonexistent.

Fig. 5.7 Mechanism of ubiquinol oxidation as deduced from structural analysis of a hydroxyquinone anion (HHDBT) and SMA binding at the Q_o site. The oxidised ISP is indicated by black circles, with the reduced ISP shown in gray. Hydrogen bonds which stabilise the enzyme-substrate complex as well as the b-position of the Rieske catalytic domain are indicated with dotted lines. *Panel 1*: Empty Q_o site with Glu272 directed out of the binding pocket. *Panel 2*: Initial stabilisation of ubiquinol by cytochrome b residues. *Panel 3*: The electron donor complex (enzyme-substrate complex) with the Rieske protein docked in the b-position. *Panel 4*: Coupled electron-proton transfer to the Rieske protein as deduced from the binding of SMA. *Panel 5*: Stabilisation of the anti-ferromagnetically coupled ubisemiquinone anion and rotational displacement of protonated Glu272 as seen for HHDBT binding. *Panel 6*: Release of the reduced and protonated Rieske protein and of the oxidised ubiquinol, accompanied by displacement of Tyr279. Steps 4 and 5 have to be interpreted as either intermediates of a sequential reaction or as transition state intermediates as proposed by the concerted mechanism hypothesis. (H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, 278, 31303-31311.)

It has been suggested that SMA and alkyl-hydroxydioxobenzothiazoles mimic either intermediates during ubiquinol oxidation, or transition state intermediates.[45] This allows for the possibility that a stable, anti-ferromagnetically coupled ubisemiquinone might be formed under some conditions. SMA binding appears to mimic the binding of a protonated ubisemiquinone as shown through panel 4 of figure 5.7, whilst HHDBT binds in its deprotonated form as an anion, as shown in panel 5. In both of these cases, however, the position of the reduced Rieske protein is stabilised by the hydrogen bond to His181. This is in agreement with electron paramagnetic resonance (EPR) analysis which indicated that the reduced Rieske

cluster is preferentially located in the b-position.[53] Glu272 can accept a proton from ubiquinol or ubisemiquinone and consequently rotate towards heme b_L, as seen for HHDBT where it is not interacting with the carbonyl group. If a stable ubisemiquinone anion is formed it is stabilised by localisation of its negative charge on the oxygen atom interacting with protonated His181. After transfer of the second electron, the product is no longer stabilised and will leave the binding pocket and give more mobility to Tyr279, thus breaking its hydrogen bond to the Rieske protein. The latter can then rotate into the binding pocket.

The yeast cytochrome bc_1 protein as described by 3CX5,[41] was used throughout the docking and is cocrystallised with SMA. The nature of SMA binding in the Q_o pocket is such that it represents *panel 4* in figure 5.7. His181 is in its protonated form so that it hydrogen bonds to the carbonyl of SMA, as shown from figure 5.5, which illustrates the interactions between the protein and SMA. The mechanistic study of the Q_o binding[45] does not discuss the importance of the water molecule in bridging the gap between SMA and Glu272 through forming the hydrogen bonding network.[41] This interaction is however observed in the 3CX5 crystal structure. The 3CX5 crystal structure was selected given that it's resolution was higher (1.90 Å) than that of comparable structures, such as that of the yeast cytochrome bc_1 complex bound with cytochrome c (resolution 2.97 Å; PDB accession code 1KYO).[54] SMA is a well documented binder at the Q_o site and a known inhibitor of the bc_1 complex, thus performing molecular docking using 3CX5 provided a means for docking compounds in their bioactive conformations, providing real and accurate representations as to how the compounds would bind.

5.2.2 Developing a Docking Protocol

The first priority was to develop a suitable docking protocol which could reproduce the docking pose of the native ligand. If the docking protocol could successfully reproduce the experimental binding mode of the native/cocrystallised ligand, then this could be applied to predict the binding pose of other ligands at the active site.[1]

GOLD 5.0.1[13] was used to develop a protocol which could reproduce the binding pose of SMA in the Q_o active site, with one of the advantages of GOLD 5.0.1 being that it could incorporate water molecules into its docking calculations.[55] This was particularly useful to model the key H-bonding network observed in the Q_o active site. The yeast cytochrome bc_1 complex (3CX5)[41] was first downloaded from the PDB[15] library, and in order to reduce the computational time required to complete the docking calculations, redundant (with regard to the docking) areas of the protein were removed. Using the molecular visualisation program PyMOL,[56] the native ligand SMA was selected and only the amino acids within 20 Å around it selected. This reduced the protein to a smaller subunit which contained only the structural information required for docking at the Q_o active site.

The docking protocol was developed using the reduced protein, with the aim being to emulate the native binding pose of SMA in the active site. A calculated docking pose was considered successful if it had an RMSD value of less than 2 Å[1] when compared to the cocrystallised ligand, ultimately providing validation for the protocol. The GOLD 5.0.1[13] wizard was used to develop the Q_o docking protocol, the full details for which can be found in the *'Q_o Docking Protocol'* as described in the Experimental Chapter. In this protocol the SMA ligand was used to define the

Q_o site ready for docking, with the cocrystallised water molecule routinely incorporated into docking to bridge the Glu272 interaction. Additionally, the tautomeric state of the His181 amino acid had to be inverted such that the hydrogen atom of the imidazole was in line with the carbonyl group of SMA, so that the required H-bond could be formed, as per the literature.[39, 45] (Any alterations to the standard protocol are discussed where appropriate.)

5.2.2.1 Docking of Stigmatellin

The native SMA ligand of the 3CX5 crystal structure was removed from Q_o and then docked back into the active site using the GOLDScore[11, 12] fitness function as described in the *'Q_o Docking Protocol'*. The resulting poses were then rescored using ChemScore.[10, 29, 30] This scoring and rescoring approach has previously proven useful in not only identifying native binding poses, but also in virtual screening.[1] Ten GA runs were performed so that a range of poses could be collected. Table 5.2 illustrates the GOLDScore and ChemScore values across these ten poses, as well as the RMSD values when they were compared to the native ligand.

Table. 5.2 GOLDScore, ChemScore and RMSD values for the ten SMA docking poses at Q_o.

Pose ID	GOLDScore	ChemScore	RMSD
1	111.6	57.2	0.64
2	107.3	57.4	0.64
3	84.4	57.7	0.86
4	77.5	52.4	4.41
5	95.6	54.7	1.00
6	106.5	56.3	0.51
7	72.9	50.8	1.02
8	95.4	50.9	1.29
9	94.9	54.6	1.00
10	76.7	47.8	1.35
Highest	111.6	57.7	4.41
Lowest	72.9	47.8	0.51
Average	92.3	54.0	1.27
SD	13.8	3.4	1.14

Given that a docking run is considered successful if it can reproduce the binding pose of the native ligand with an RMSD value of less than 2 Å,[1] table 5.2 shows that only one of the ten poses failed to meet this criterion. With this in mind it should be remembered that by its very nature, the docking GA incorporates an element of randomness,[12, 57-59] so the same solutions will not always be found. The smallest RMSD value reported was only 0.51 Å, with the largest being 4.41 Å. This gave an average RMSD value across the ten poses of only 1.27 Å, and an SD of 1.14. The average result is very encouraging, and is perhaps a more fair way of assessing the performance of the docking, as oppose to considering any one result in isolation. Taking an average result allows for consideration of all the poses generated, not just the most suitable identified through visual inspection. Removing the highest RMSD value for pose 4 which fell outside the cut off of 2 Å improved the average RMSD value further to 0.92 Å, with an SD of 0.29.

The GOLDScore and ChemScore values can be considered means of assessing how well a compound binds to a particular site, with higher values suggesting better binding. The average GOLDScore value across the ten poses was 92.3 ± 13.8, with the highest individual value reported as 111.6. It is interesting to note that the pose which had the lowest RMSD value (4.41 Å), also had one of the lowest GOLDScores (77.5). It also had a low ChemScore value of 52.4, slightly lower than the average across all poses of 54.0 ± 3.4. Therefore, given that the better poses (i.e. those with smaller RMSD values) also observed higher fitness function scores, this is highly encouraging and suggests that the docking protocol was suitable for molecular docking at the Q_o binding site. Figure 5.8 represents the binding pose of one of the solutions that was generated from the docking of SMA. It had a high GOLDScore of 107.3, a ChemScore of 57.4, and an RMSD of 0.64 Å. The results of SMA docking could be used for comparison between the scores of other docked compounds, in order to draw comment on their inhibition potential.

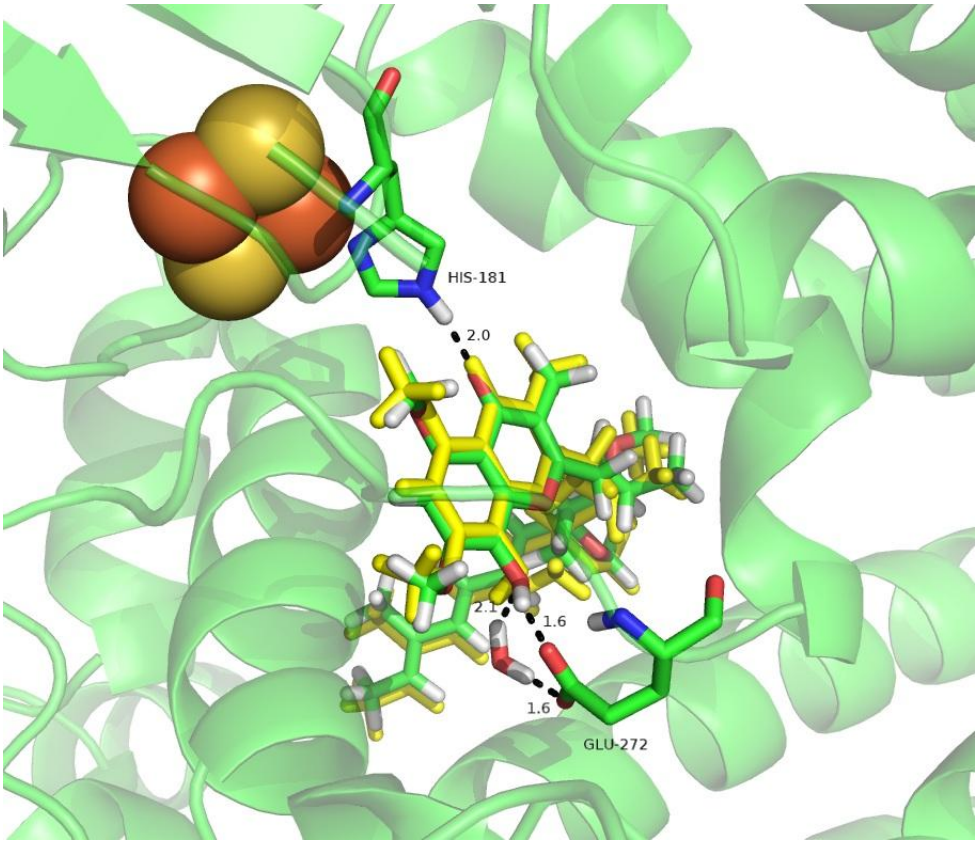


Fig. 5.8 Docking pose of SMA (shown in green) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The two key interactions, His181 and the water mediate Glu272 H-bond network are clearly illustrated. The native binding pose of SMA is shown in yellow. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

The docking of SMA was repeated using the *'Q_o Docking Protocol'*, only this time the bridging water molecule was omitted to see what affect (if any) this would have on the results. Ten docking poses were collected, the scores for which are shown in table 5.3.

Table. 5.3 GOLDScore, ChemScore and RMSD values for the ten SMA docking poses at Q_o, when the crystallographic water molecule was removed.

Pose ID	GOLDScore	ChemScore	RMSD
1	70.3	55.4	0.75
2	100.8	55.9	0.69
3	64.7	46.4	1.22
4	63.6	51.2	0.81
5	81.5	47.1	1.37
6	50.7	42.7	3.21
7	69.1	47.4	1.23
8	79.9	54.0	0.33
9	94.4	54.6	0.87
10	103.5	57.4	0.56
Highest	103.5	57.4	3.21
Lowest	50.7	42.7	0.33
Average	77.8	51.2	1.10
SD	17.4	5.0	0.81

The ten docking solutions all produced reasonable poses, with an average RMSD value compared to SMA of 1.1 Å. An example of one solution is shown in figure 5.9. The average GOLDScore and ChemScore values were 77.8 ± 17.4 and 51.2 ± 5.0 respectively, lower than the solutions from docking with the water present. This reduction in the fitness functions is most likely due to the way in which the scores are calculated. The absence of the water molecules resulted in the loss of additional H-bonds. As these bonds contribute to the score, their omission resulted in smaller scores.

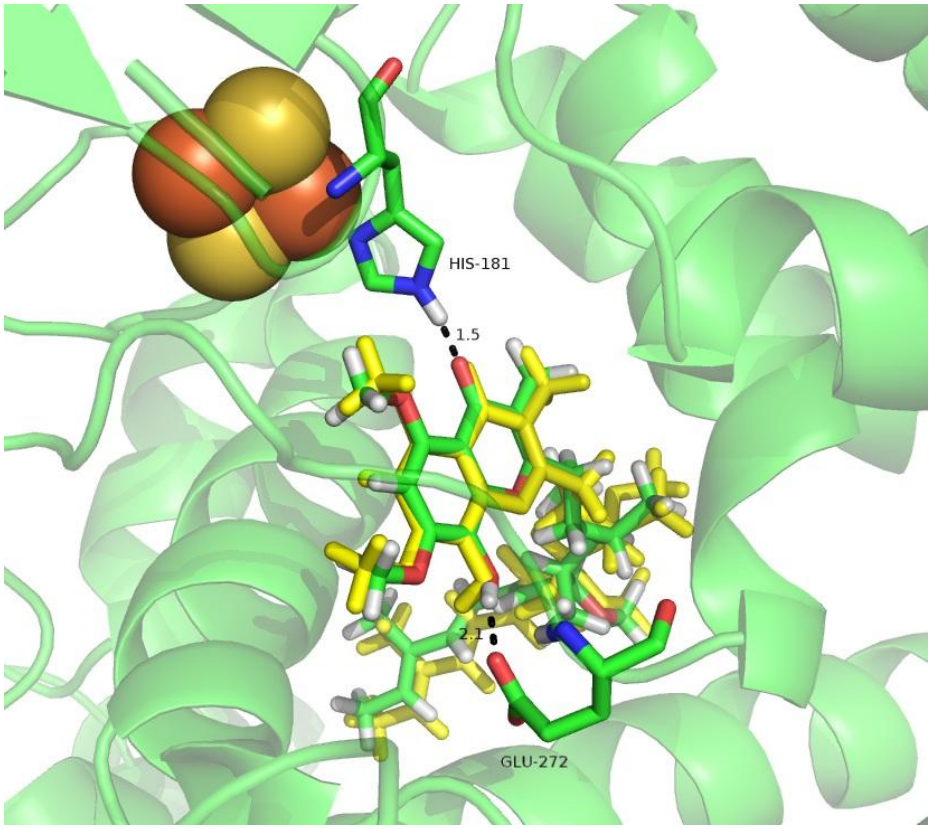


Fig. 5.9 Docking pose of SMA (shown in green) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5) when water is omitted from the docking. The His181 interaction is observed along with a partial Glu272 H-bond network. The native binding pose of SMA is shown in yellow. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Not only do the fitness scores suffer when the water molecule is removed, but it must also be remembered that the water mediated interaction is supported by precedent in the literature.[39] Inclusion of the water resulted in higher fitness scores and ultimately produces more meaningful solutions. Given that GOLD 5.0.1[13] readily has the ability to include this water molecule in its docking runs, as well as the observed improvement to the fitness scores, there seems no reason to exclude it. Thus, all docking from here on was performed with the water molecule included.

5.2.2.2 Docking of Atovaquone

ATOV is a class I inhibitor,[32] competing with ubiquinol at the Q_o site.[60] It is currently the only drug in clinical use which targets *Pf*bc_1,[33, 61, 62] acting by collapsing the mitochondrial membrane potential. However, it has yet to be cocrystallised with the bc_1 complex, which made it an attractive compound for a molecular docking study. High levels of resistance have been observed which are correlated to a number of point mutations in cytochrome b. ATOV binds to the Q_o site when the soluble domain of the Rieske protein is proximal to cytochrome b, and interacts directly with the ISP. This prevents mobilisation to cytochrome c_1, and consequently impairs the mitochondrial transmembrane potential.[61, 63-65] The yeast crystal structure of the cytochrome bc_1 complex was therefore suitable for docking, owing to its high sequence homology with the *plasmodium*.[42]

Spartan '08[66] was first used to construct ATOV *in silico* and then minimised using the *'Energy Minimisation Protocol'* as described in the Experimental Chapter. The equilibrium geometry calculation was performed using the molecular mechanics, MMFF level of theory,[67] which was sufficient for the purpose of this application. The *'Q_o Docking Protocol'* was used to perform the molecular docking of ATOV into the Q_o site, albeit with several modifications. Given that it was not simply the native ligand being docked back into the active site, the number of GA runs was increased from 10 to 25. To some degree the Q_o site represents an SMA shaped pocket, so the increased number of GA runs allowed for a bigger range of solutions to be explored, capturing more diversity in the poses. From these the most appropriate solutions were identified so that only those which were in line with literature precedent were considered. The HOH7187 water molecule was again

incorporated into the docking calculation to mediate the interaction between the carbonyl of the quinone ring and Glu272, only this time it was allowed to translate from its original position within a radius of 2 Å, to better bridge the gap. Additionally, in order to ensure ATOV docked into the Q_o site as appose to the hydrophobic pocket, a constraint was applied so that poses would be biased towards forming H-bond interactions with His181. Though SMA docking showed the formation of a H-bond between the hydrogen of the His181 imidazole and the carbonyl group of SMA, for ATOV it has not been fully established which group contributes the hydrogen atom in the putative H-bond between His181 and the naphthoquinone hydroxyl of ATOV.[60] Thus, the His181 residue was taken to be in its imidazolate ionisation state. The resulting 25 solutions consisted of varying poses, with only four resembling the orientation suggested by the literature, shown in figure 5.10.[32, 64]

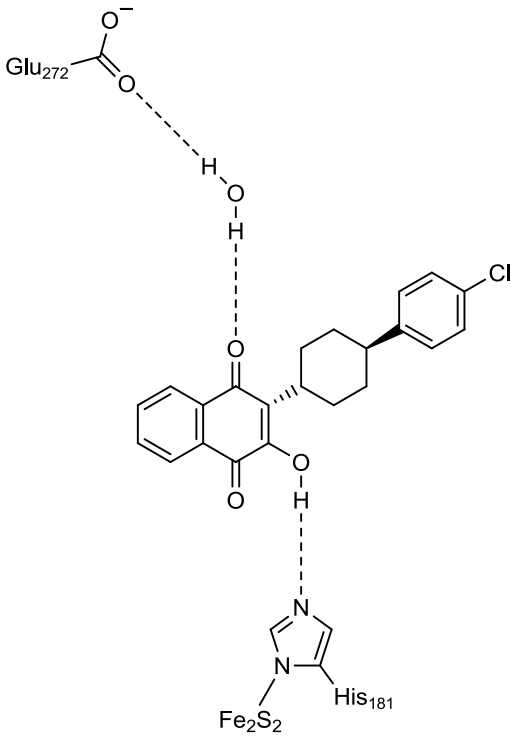


Fig. 5.10 H-bond interactions of atovaquone within the Q_o site. (J. J. Kessl, B. B. Lange, T. Merbitz-Zahradnik, K. Zwicker, P. Hill, B. Meunier, H. Palsdottir, C. Hunte, S. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2003, **278**, 31312-31318.)

The other 21 poses had the naphthoquinone hydroxyl head group in the hydrophobic pocket, and given that ATOV is a known Q_o inhibitor, we were only concerned with its docked orientation in the binding site, and not its resulting fitness scores. Therefore, only the GOLDScore fitness function was used as no comparison between ATOV and other molecules was necessary.

The four poses in an orientation similar to that of figure 5.10 all observed the two crucial H-bond interactions, and had an average GOLDScore of 28.9 ± 4.1 (table 5.4). Though this is less than the average GOLDScore of SMA docking (92.3 ± 13.8), it should be remembered that hydrophobic and van der Waals interactions contribute greatly to the overall GOLDScore, and thus, considering SMA has a long hydrophobic side chain, many more such interactions are formed, resulting in a higher score. The aryl side chain of ATOV is much more rigid than that of SMA, and forms fewer such interactions.

Table. 5.4 GOLDScore values for the four ATOV docking poses at Q_o which resembled figure 5.10.

Pose ID	GOLDScore
1	29.0
2	32.5
3	23.2
4	31.0
Highest	32.5
Lowest	23.2
Average	28.9
SD	4.1

This docking study was used to assist in our understanding as to how ATOV binds at the Q_o site and elicits its response. It also allowed for comment to be drawn as to how ATOV activity can be lost through mutation. Resistance to ATOV is associated with a missense point mutation in cytochrome b of the Q_o pocket.[32] The mutation occurs at position 268, exchanging tyrosine for a serine (Y268S), or less frequently,

asparagine (Y268N).[68-73] Ordinarily, position Y268 in cytochrome b is highly conserved across all phyla, with the Pro_{260}-Glu_{261}-Trp_{262}-Tyr_{263} ('PEWY') region located in the ef loop of the Q_o site. The side chain of the Y268 residue most likely forms a stabilising hydrophobic interaction with bound ubiquinol, and studies have suggested that ATOV is similarly stabilised.[60] For the yeast cytochrome bc_1 complex, the introduction of the Y268S mutation (Y268S) resulted in an increased IC_{50} value, from 60 nM to >4000 nM.[74] This result validates the significance of our docking at the Q_o site of 3CX5, as ATOV clearly inhibits the yeast cytochrome bc_1 complex, making a mechanistic study attractive. As no crystal structure of the parasite cytochrome bc_1 complex currently exists, yeast is therefore a useful surrogate.

The docking study was an updated version of previous work which had been performed,[75] and the results were analysed to investigate the orientation of ATOV in the Q_o active site. ATOV was found to bind in a manner very similar to SMA, with the hydroxyl moiety of the hydroxynaphthoquinone ring forming a hydrogen bond to the nitrogen of the imidazolate group of His181 in the Rieske ISP (lowering the redox potential of the [2Fe2S]). A second H-bond was formed between ATOV's hydroxynaphthoquinone carbonyl group and the carboxylate of the cytochrome b ef loop residue Glu272, via the bridging water molecule. These interactions are shown in figure 5.11. This work has been published in *The Journal of Biological Chemistry* as part of the paper titled *'Cytochrome b Mutation Y268S Conferring Atovaquone Resistance Phenotype in Malaria Parasite Results in Reduced Parasite bc1 Catalytic Turnover and Protein Expression'*.[76]

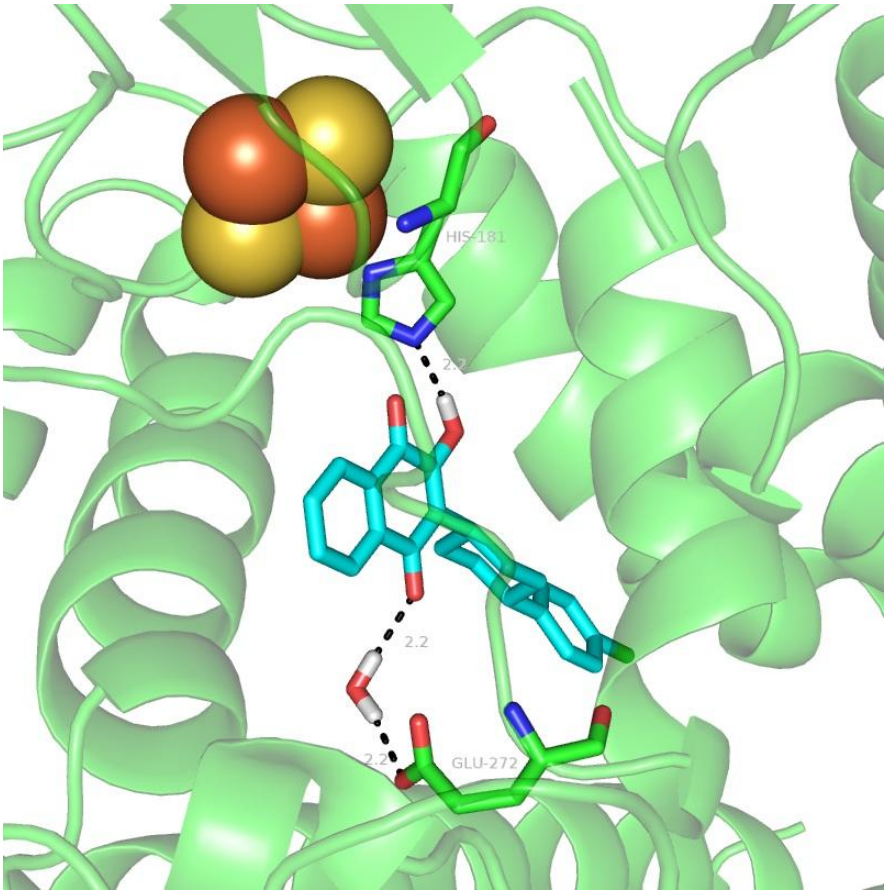


Fig. 5.11 Docking pose of ATOV (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The two key interactions, His181 and the water mediate Glu272 H-bond network are clearly illustrated. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Additionally, the chlorophenyl ring of ATOV appears to sit in a hydrophobic pocket within cytochrome b, formed from the side chains of Phe121, Phe278, Ile125 and Ile299. These can be seen in figure 5.12.

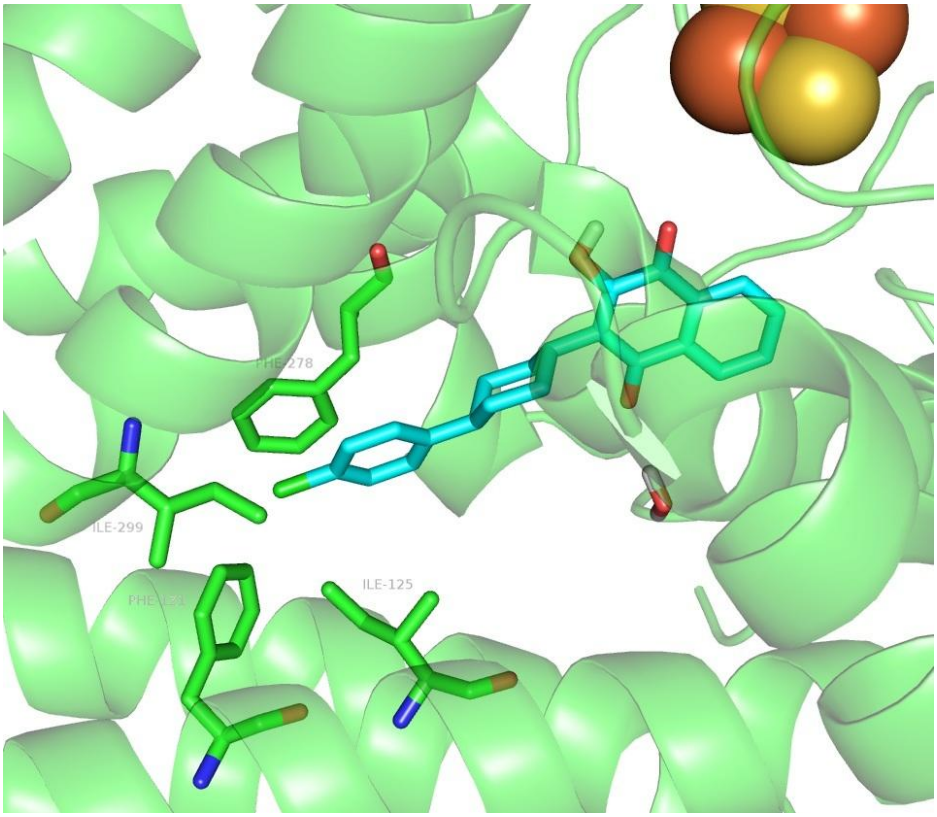


Fig. 5.12 Docking pose of ATOV (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The chlorophenyl ring of ATOV sits in a hydrophobic pocket formed by the side chains of Phe121, Phe278, Ile125 and Ile299. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange).

Also, the amino acid Leu275 is predicted to form a stabilising hydrophobic contact with ATOV's cyclohexyl moiety, as shown in figure 5.13. All of these interactions are likely to occur in the corresponding residues of *Pf*bc_1.

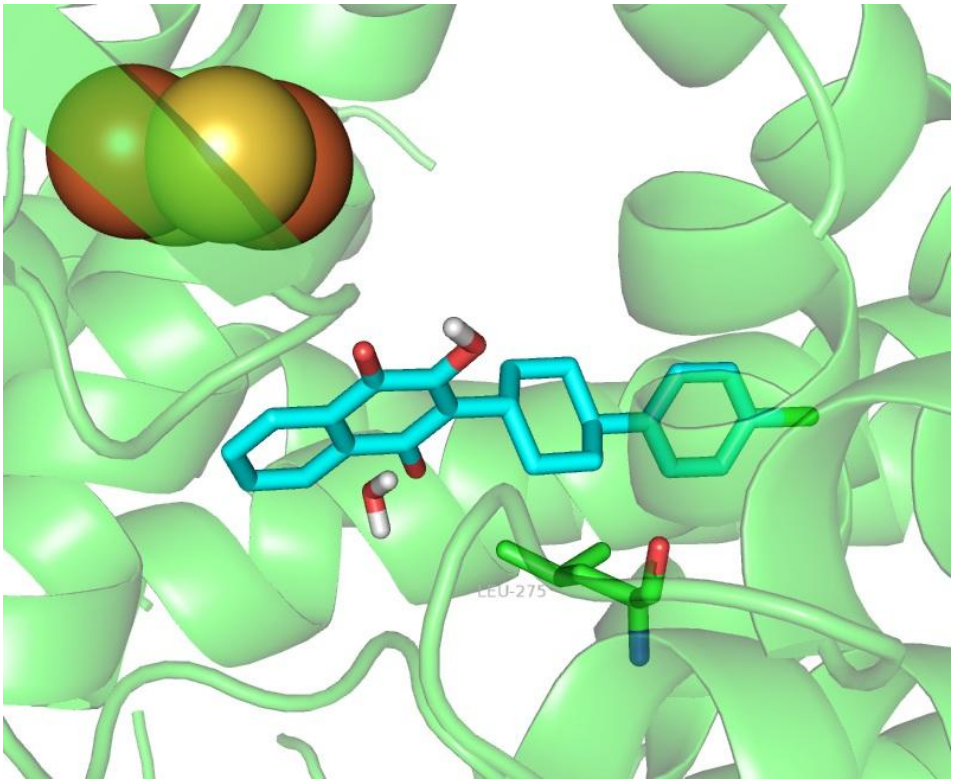


Fig. 5.13 Docking pose of ATOV (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The cyclohexyl moiety forms a stabilising hydrophobic contact with Leu275. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange).

Y268 (*Plasmodium* numbering) is located near the PEWY region of the ef helix and is highly conserved. In the yeast protein, the PEWY region corresponds to the amino acid residues Pro_{271}-Glu_{272}-Trp_{273}-Tyr_{274}. Subsequently, the Tyr268 residue in the *Plasmodium* corresponds to Tyr279 in the yeast. It has been suggested that the tyrosyl side chain of this residue participates in the positioning of Q_o bound ubiquinol, and may also therefore contribute to stabilising hydrophobic interactions with the naphthoquinone group of ATOV.[45, 75] Studies with mutant forms of bc_1 from the bacterium *Rhodobacter sphaeroides*, indicate that an aromatic or large hydrophobic side chain residue is required at this position within the ef loop for effective catalytic activity.[77] Therefore, the point mutation of this amino acid to either a serine or asparagine may result in loss of activity. Examination of the crystal structure of avian bc_1 suggests that the hydroxyl group of this tyrosyl side chain may

form a H-bond association with His181 of the Rieske protein (figure 5.14).[78] In addition, mutations of the equivalent residue in man (Tyr279) have been linked to a variety of mitochondrial disorders.[79, 80] As a whole, this docking study illustrates the importance of several amino acid residues in the binding of ATOV, and shows the drastic effect even a single point mutations can have on activity.

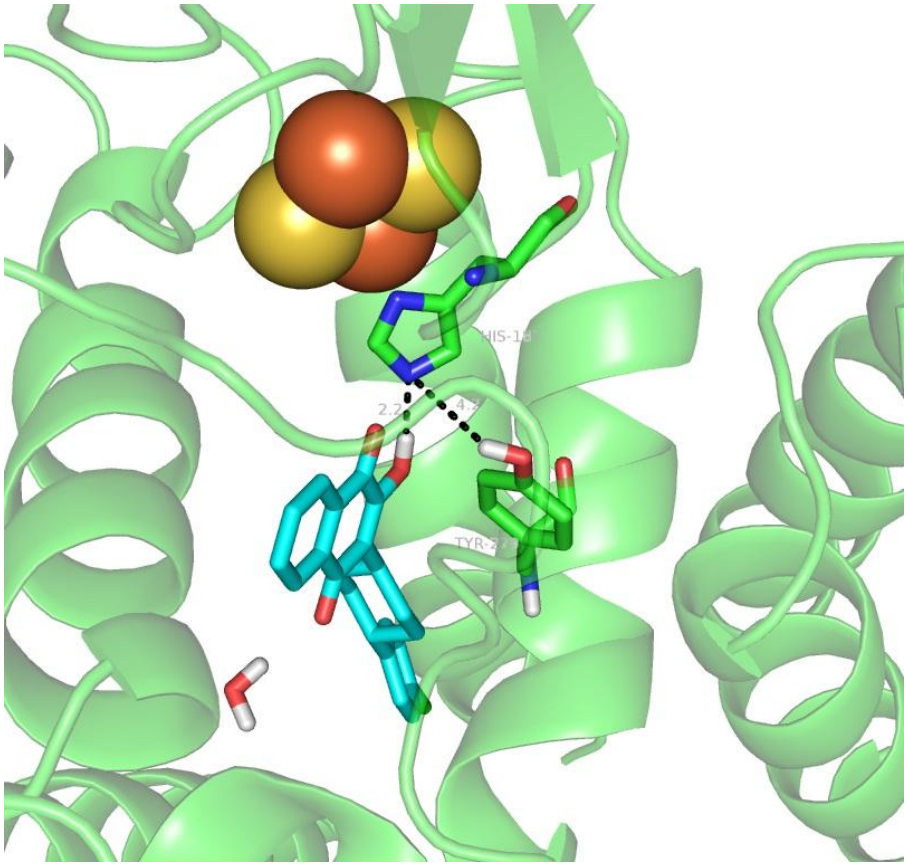


Fig. 5.14 Docking pose of ATOV (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The hydroxyl group of Tyr279 forms a stabilising H-bond association with His181 of the Rieske protein. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

5.2.2.3 Docking of Quinolone Esters

Based on the key interactions associated with activity at the Q_o active site (His181 and Glu272), it can be deduced that a polar head group is required in order to form the H-bond associations, together with an alkyl or aryl side chain with which to

probe the cylindrical hydrophobic pocket. This rationalisation has been supported by work which led to structures such as the quinolone ester in figure 5.15.[81] However, whilst this core structure made for an interesting chemotype, the long flexible perfluorinated alkyl side chain was less than ideal in terms of its potential as a drug lead.

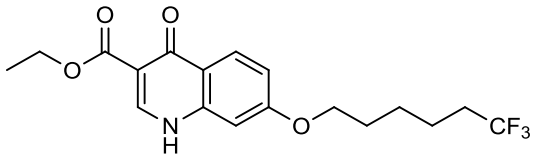


Fig. 5.15 Quinolone ester.

A potential advantage of the quinolone template is that when suitably substituted, it may yield compounds with the ability to chelate haem through π-stacking, in a similar manner to the 4-aminoquinolines.[82] This could allow for a single drug capable of attacking the parasite by two separate mechanisms, a feature that may hinder the development of resistance. From this, work within the antimalarial group at Liverpool aimed to identify suitable alternatives to the compound in figure 5.15 which had the potential to drive forward the synthesis of quinolone esters.[42] In all, twenty novel quinolones were developed as part of two parallel series based on the single core template shown in figure 5.16.[83] The Gould-Jacobs methodology[84-86] was employed to synthesise a number of 6- and 7-substituted aryl quinolones. These were then tested *in vitro* against cultured wild type 3D7 *P. falciparum* malarial parasites (see *'Whole Cell Growth Inhibition Assay (3D7) Protocol'* in Experimental Chapter). Activity values spanned a wide range of activities, from sub nM (0.46 nM) to several μM (>10 μM)

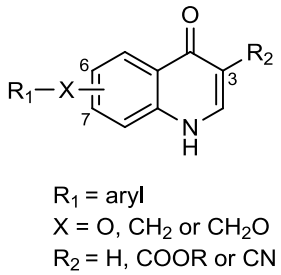
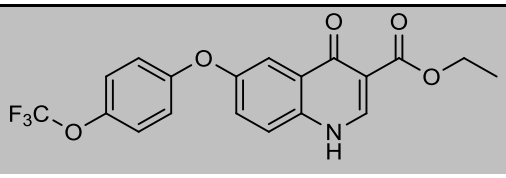
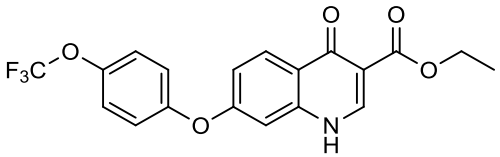


Fig. 5.16 Core quinolone template.

Whilst the synthesis, testing and analysis of the entire quinolone series does not form part of this thesis, molecular modelling was employed in an attempt to rationalise the key observation which emerged from the SAR study. It was found that in almost all cases, the 7-aryl substituted compounds had superior activity to their 6-substituted counterparts, with one 7-substituted compound some 300 times more potent than the 6-substituted analogue. Given that this trend was true across most of the series, it was thought that by studying their binding modes in the Q_o site, insight may be garnered as to why the activity of the 6- series was drastically reduced.

Docking was therefore performed for two exemplary molecules from this data that observed this trend. Compounds Xa and Xb are shown in table 5.5, with their respective IC_{50} values against 3D7 reported as 229.8 nM and 5.3 nM. Compounds Xa and Xb were first drawn in Spartan '08[66] and energy minimised according to the *'Energy Minimisation Protocol'*.

Table. 5.5 Quinolone ester compounds Xa and Xb.

Compound ID	Structure	IC_{50} 3D7 (nM)
Xa		229.8
Xb		5.3

Docking was then performed using the *'Q_o Docking Protocol'* with the GOLDScore fitness function. From the ten GA runs performed some interesting observations were found. Eight out of the ten poses for Xb docked in a very similar manner to SMA, with an average GOLDScore value of 62.8 ± 1.4. This may go some way to rationalising its potent *in vivo* activity. One solution is shown in figure 5.17, and had a GOLDScore value of 62.8 (coincidentally the same as the average). Of particular importance in this pose was the strong H-bond between His181 and the carbonyl group of the quinolone, as this prevents proton transfer between ISP and ubiquinol, and thus inhibition of subsequent electron transfer to ISP. Additionally, the aryl side chain of Xb rests in a hydrophobic pocket within cytochrome b, comprised of Phe121, Phe278, Ile125 and Ile299. This is very similar to ATOV binding. Also, though it is more distant, the NH group of the quinolone could potentially form a H-bond with Glu272 under the right circumstances.

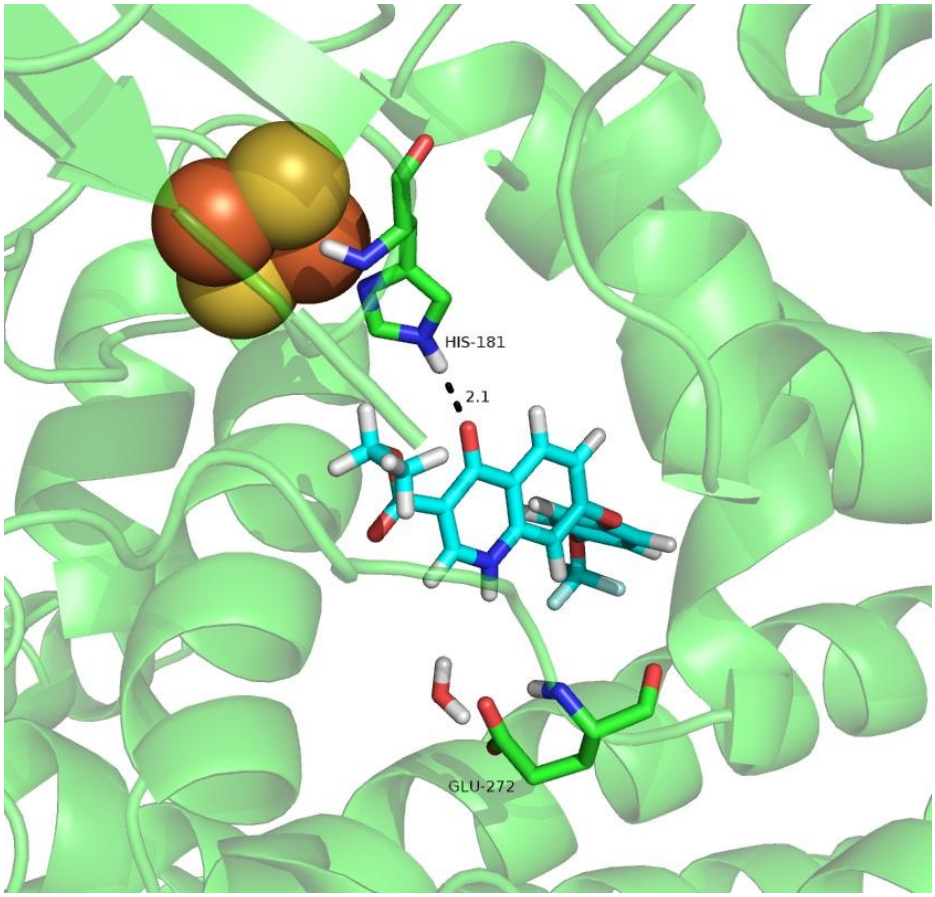


Fig. 5.17 Docking pose of Xb (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). His181 is shown forming a strong H-bond with the carbonyl oxygen of the quinolone. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

The ten docking solutions for Xa all produced similar poses, and had an average GOLDScore value of 65.1 ± 3.3. However, the orientation of the quinolone head group was flipped in the active site when compared with Xb, such that the crucial His181 H-bond interaction could not be formed, as shown by figure 5.18. Despite this, the water mediated Glu272 interaction was readily observed with the carbonyl group of the quinolone, but an unfavourable interaction existed between the NH of His181, and the NH of the quinolone. Given the nature of activity at the Q_o site, this observation may therefore account for the drastically reduced activity of the 6-aryl substituted series, as they do not bind strongly enough to prevent ubiquinol binding, and thus inhibit the ETC. Whilst the GOLDScores may be comparable between Xa

and Xb, all they really tell us is that the sum of the *in silico* interactions suggest similar binding affinities, but this appears to be unrelated to their efficacies. The orientation of the molecules in the active site therefore has a dramatic effect on *in vitro* activity.

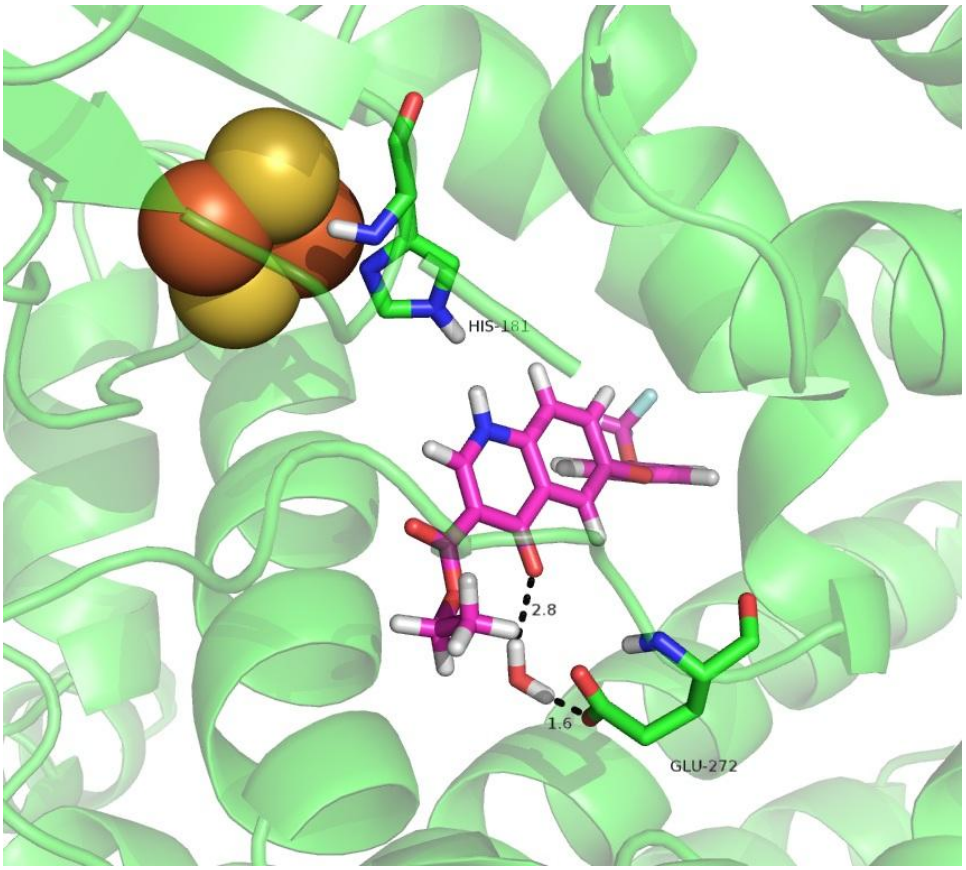


Fig. 5.18 Docking pose of Xa (shown in pink) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). Unfavourable His181 and NH interaction observed, though water mediated H-bond network readily formed between the carbonyl group and Glu272. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

This binding behaviour may be explained by consideration of the steric bulk of the compounds. Initial docking showed both Xa and Xb to fit nicely into the Q_o pocket as shown, with the aryl side chains moving out into the hydrophobic pocket. With this in mind, it was decided to bias the quinolone head group of the 6-substituted compound into an orientation which resembled that of the 7-substituted compound, to see how it would behave. To do this the docking was repeated for both

compounds with constraints applied similar to those used in ATOV docking, so that poses would be biased towards forming H-bond interactions with His181 and Glu272.

Compound Xb docked as before for all ten solutions, this time giving an average GOLDScore value of 42.2 ± 2.4 (value different to before owing to constraints affect on scoring). The ten poses for Xa also observed the same orientation, with an average GOLDScore of 36.9 ± 2.4. However, all were noticeably strained in their positioning, giving poses such as that in figure 5.19 (GOLDScore of 33.9). As can be seen, the carbonyl group of the quinolone forms a H-bond with His181, and the water mediated Glu272 interaction also has the potential to be observed. Despite this, when the pose was compared with that of Xb, concerns with regard to its validity were raised.

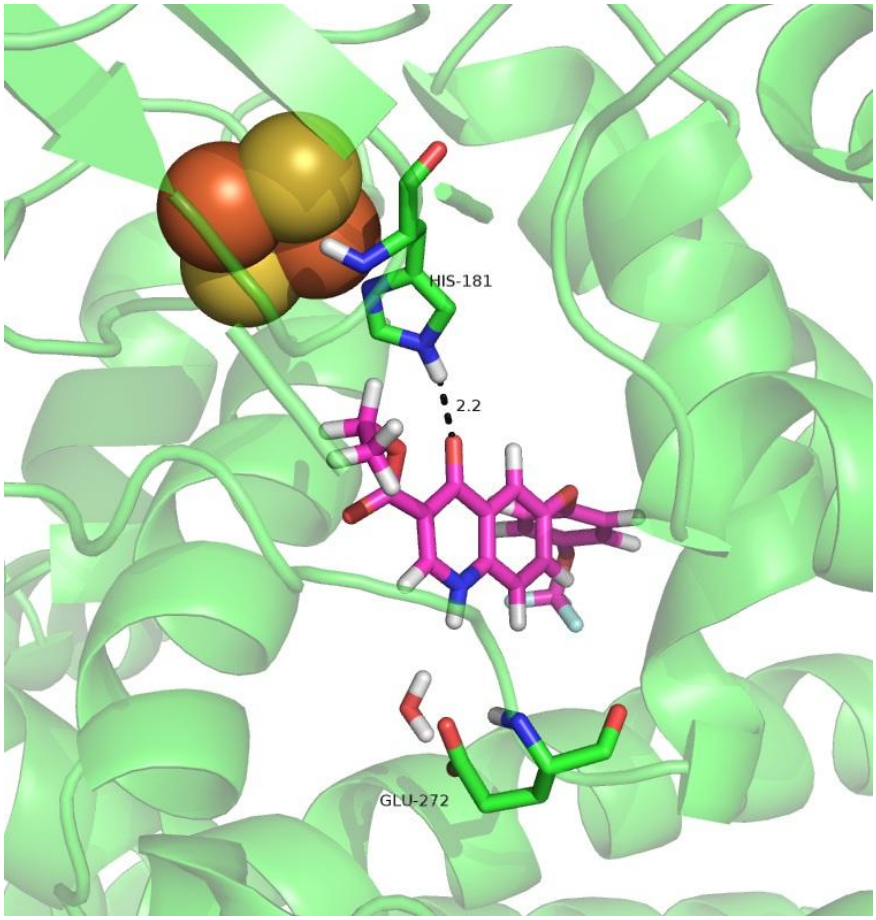


Fig. 5.19 Docking solution of Xa (shown in pink) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5) when constraints were applied. The His181 and quinolone carbonyl H-bond is clearly observed, as is the potential for the Glu272 interaction. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Figure 5.20 illustrates the strain of Xa in the Q_o site when compared with one of the Xb solutions (GOLDScore of 45.4). The figure shows the surface of the protein as a wire mesh, making it easier to spot the steric clashes. Xb (shown in blue) rests nicely in the pocket with no clashes to neighbouring amino acids. However, the methoxy linker of the aryl side chain of Xa (shown in pink) appears to sit very closely to the sulphur atom of Met295. It is highly unfavourable for these two electronegative atoms to be so close to one another, owing to the repulsive nature of their respective lone pairs. Additionally, the steric strain of their close proximity would also make the positioning of Xa in this orientation highly unlikely.

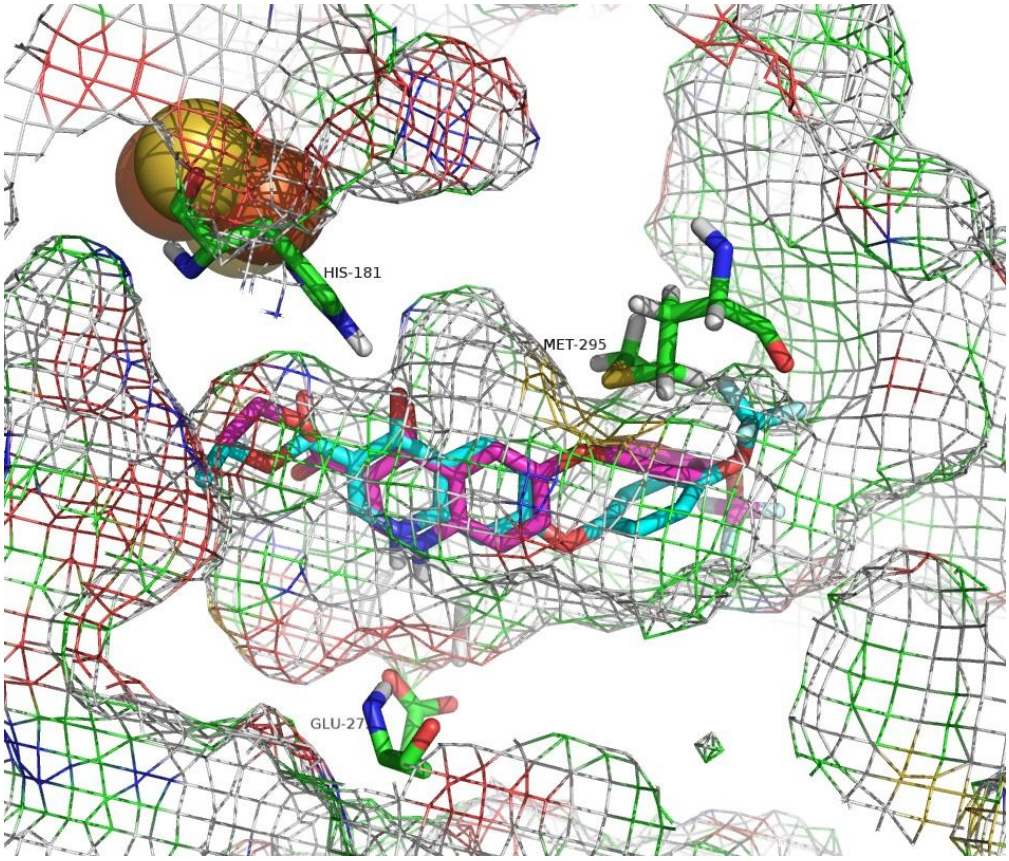


Fig. 5.20 Docking solutions of Xa (shown in pink) and Xb (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5) when constraints were applied. The methoxy linker of Xa is shown to have an unfavourable steric clash with Met295. The protein surface is represented as a wire mesh, with the [2Fe2S] cluster of the Rieske protein shown as spheres (sulphur: gold, iron: orange).

This study showed that the 7-substituted analogues of the quinolone series bound similar to SMA, with the quinolone head group mimicking the transition state semiquinone observed during the catalytic turnover of the bc_1 complex, providing a H-bond donor and acceptor site.[87] For the 6-substituted analogues however, these sites were not available, as the quinolone head group preferentially flipped during docking, owing to steric and repulsive interactions. Given that the His181 interaction was no longer observed, this may provide an explanation as to why the 6-substituted quinolones had dramatically lower activity values compared to their 7-substituted counterparts. Quite simply, the molecules were unable to fit in the Q_o site in a manner which promoted inhibition. Observations from this docking study were reported as part of a publication in *MedChemComm* entitled *'The development*

of quinolone esters as novel antimalarial agents targeting the Plasmodium falciparum bc1 protein complex'.[83]

5.2.2.4 Docking of Lead Quinolone Compound

At Liverpool research is continually ongoing to identify new lead structures for drug development that are cost effective, and capable of inhibiting several targets in an attempt to circumvent developing resistance. One such compound which is currently leading the way for a new generation of antimalarial quinolone compounds is SL-2-25, shown in figure 5.21. SL-2-25 has shown to be a promising candidate for the treatment and prophylaxis of uncomplicated malaria.

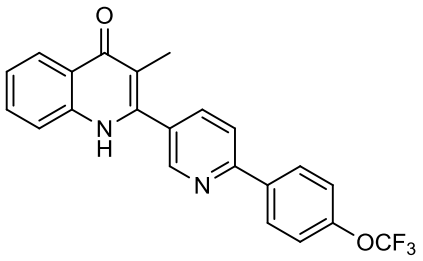


Fig. 5.21 Compound SL-2-25.

It is comprised of a novel, pyridine containing aryl side chain in the 2-position of the quinolone chemotype, and has been found to have a duel mechanism of action against two of the respiratory enzymes in the electron transport chain. The activity values for this compound against several assays are reported in table 5.6. Whole cell inhibition of SL-2-25 against the 3D7 parasite strain was 54 nM, with targeted activity against *Pf*NDH2 and *Pf*bc_1 of 14.6 nM and 15.1 nM respectively. This gives a selectivity value between the two of 1.03, indicating the strong duel inhibition capabilities of the compound. Additionally, bovine bc_1 inhibition was reported as 887 nM. As has previously been discussed, strong inhibition of bovine bc_1 is associated with cardiotoxicity in mammals,[88] thus weak inhibition of bovine bc_1 is

highly desirable. This new generation of antimalarials therefore has the potential to overcome the limitations of existing treatment options such as ATOV, as it may circumvent developing resistance mechanisms.

Table. 5.6 SL-2-25 activity values.

Compound	3D7 IC_{50}	*Pf*NDH2	*Pf*bc_1	Bovine bc_1
SL-2-25	54 ± 6 nM	14.6 nM	15.1 nM	887 nM

Molecular docking was employed to rationalise the low nM activity of SL-2-25 against *Pf*bc_1, by docking it into the Q_o site. Once again, the yeast bc_1 complex was used as a surrogate for the parasite, which though not identical, does share an overall 40% sequence homology and is highly conserved across the Q_o region.[42] SL-2-25 was constructed and energy minimised using the *'Energy Minimisation Protocol'*, with docking performed using the *'Q_o Docking Protocol'*, only with 25 GA runs performed as oppose to 10. The crystallographic water molecule present in the Q_o site was incorporated into the docking run to mediate the Glu272 interaction, and allowed to translate from its original position within a radius of 2 Å. Constraints were applied to only give solutions which formed H-bond interactions with His181 and Glu272.

All 25 solutions showed SL-2-25 in the same orientation, giving average GOLDScore and ChemScore values of 53.7 ± 2.2 and 30.7 ± 1.0 respectively. One solution is shown in figure 5.22 (GOLDScore and ChemScore values of 46.9 and 31.9 respectively). SL-2-25 occupies a position within the Q_o site of cytochrome b similar to SMA. A strong H-bond is observed between the carbonyl of the quinolone head group and the imidazole ring of Rieske protein residue His181. A water bridged H-bond is also observed between the quinolone NH group and Glu272 of the

PEWY motif. Weaker hydrophobic and van der Waals interactions are also observed.

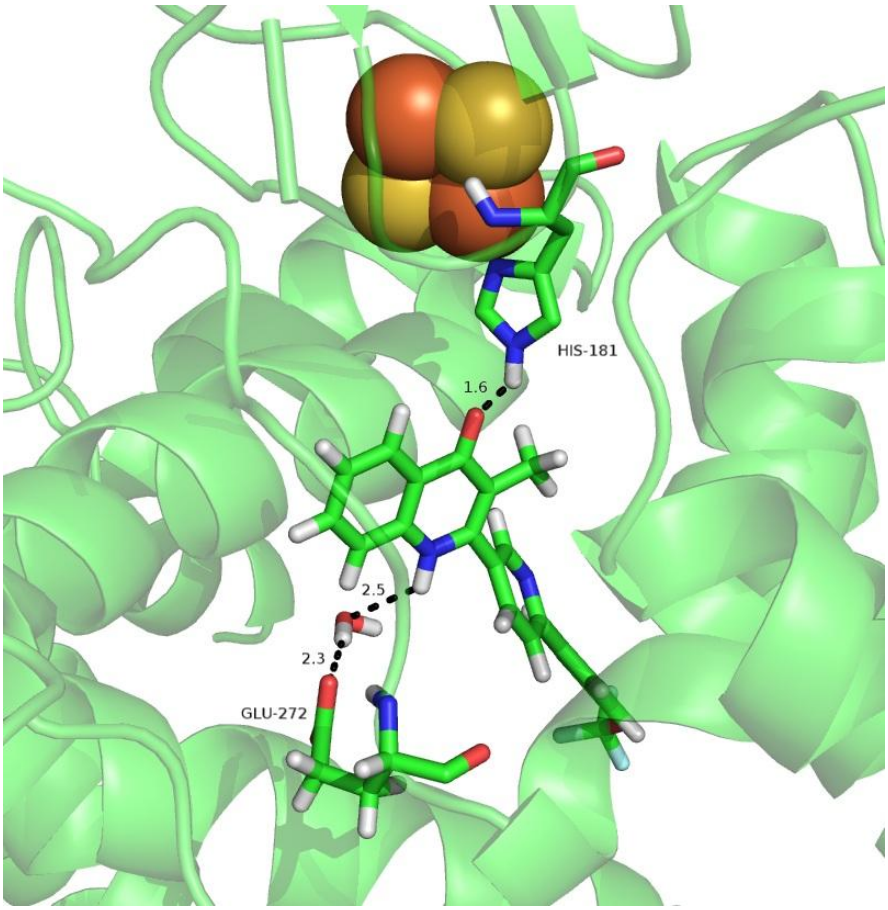


Fig. 5.22 Docking solution of SL-2-25 (shown in green) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5) when constraints were applied. The His181 and quinolone carbonyl H-bond is clearly observed, as is the Glu272 water mediate interaction with the NH group of SL-2-25. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

The observations from SL-2-25 docking formed part of a publication in the '*Proceedings of the National Academy of Sciences'* journal entitled *'Generation of quinolone antimalarials targeting the Plasmodium falciparum mitochondrial respiratory chain for the treatment and prophylaxis of malaria'*.[89]

The docking of ATOV and the quinolone compounds showed that the His181 and Glu272 H-bonds are consistently observed in molecules which are highly active

towards the bc_1 complex. Though this had already been postulated in the literature based on previous work,[39] the significance of this observation is that it may drive forward future drug discovery efforts, as compounds can be pre-screened to assess their potential as binders at the Q_o site.

5.2.3 The Q_i Site

Similar to the Q_o site, the Q_i site is also a promising target for antimalarial drug design. Compounds which target this site are referred to as class II inhibitors, blocking the electron transfer path from heme b_H to quinone.[39, 40] However, the use of class II inhibitors has thus far been limited, as they are often highly toxic in mammals and other non-pathogenic organisms. The Q_i site is rigid and flat, but such is its size and shape that inhibitors can bind in a variety of ways based on the amino acid residues which line its pocket.[39] A few residues lining the wall of the Q_i pocket are thought to be involved in specific substrate binding, including His201, Ser205, Lys227 and Asp228 (bovine bc_1 complex numbering). However, it has been found that amino acid residues Trp31, Gly38, Met190, Met194, Leu197 and Ser35 provide specific interactions for the binding of Q_i inhibitor antimycin A (fig. 5.3), whilst the Leu200, His201, Ser205, Phe220, Tyr224, Lys227 and Asp228 residues are more important for ubiquinone binding. Structurally, these two groups of residues face each other on the opposite sides of the Q_i pocket.

5.2.3.1 Docking of Antimycin A

Antimycin A is a natural fungicide (fig. 5.3) and a known class II inhibitor.[39, 40] It binds to the Q_i site to inhibit the oxidation of ubiquinol, disrupting the formation of the proton gradient across the inner membrane, and thus ATP generation. Whereas

Q_o inhibitors such as SMA are capable of stabilising the conformation of the ISP extrinsic domain, Q_i inhibitors have no obvious effect on the mobility of ISP.[39]

A molecular docking study of antimycin A was undertaken to develop and validate a docking protocol suitable for docking at the Q_i site. Given that the Q_i bond interactions have already been reported for antimycin A,[39] these could be used as a reference point to emulate the results, thus validating the protocol. Antimycin A was first energy minimised using the *'Energy Minimisation Protocol'*, and then docked according to the *'Q_i Docking Protocol'*, as described in the Experimental Chapter. This protocol was fairly similar to that used in Q_o docking, but unlike Q_o, there were no active water molecules present in the Q_i binding site to mediate particular interactions. Also, the Q_i site of the yeast bc_1 protein was empty, so it was first aligned with that of the bovine bc_1 complex crystal structure (PDB accession code 1SQX)[39] using PyMOL,[56] as the bovine protein had been cocrystallised with ubiquinone in the Q_i site. When the two proteins were aligned it was possible to incorporate the native ubiquinone ligand from bovine into the empty Q_i site of yeast. Ubiquinone could then be used to define the Q_i site for docking. Prior to docking this modified yeast protein was first reduced in size as was done with the Q_o docking protein, by considering only the amino acid residues within 20 Å of the ubiquinone ligand. This reduced the computational time required to complete the docking runs by including only the essential structural information required. The number of GA runs was also increased from 10 to 25 to incorporate any potential diversity amongst the solutions.

The 25 poses had average GOLDScore and ChemScore values of 66.4 ± 8.1 and 22.8 ± 3.1 respectively. What is particularly interesting is that all of the binding poses

had very similar orientations, despite the highly flexible side chains and the many degrees of freedom antimycin A possesses. One such pose is illustrated in figure 5.23 (GOLDScore and ChemScore values of 75.3 and 26.0 respectively). The numbering of the amino acids in the protein differ to those which have been quoted previously, owing to the fact that this is the yeast protein and not the bovine. Where appropriate, the corresponding bovine residue will be reported in brackets. The expected interactions between Trp30 (Trp31), Gly37 (Gly38), Ser34 (Ser34) and Leu198 (Leu198) were all observed. However, it has also been reported that antimycin A interacts with Met190 and Met194, and neither of these residues are present in yeast. This suggests that perhaps the Q_i site is not as highly conserved as that of Q_o. Additionally, interactions were observed with Phe225 (Phe220), Lys228 (lys227) and Asp229 (Asp228), forming a nice pocket for the amide group on the benzene ring to sit in. Most of the interactions appeared to be either hydrophobic in nature, or simple van der Waals contacts. These observations support the validity of the *'Q_i Docking Protocol'*.

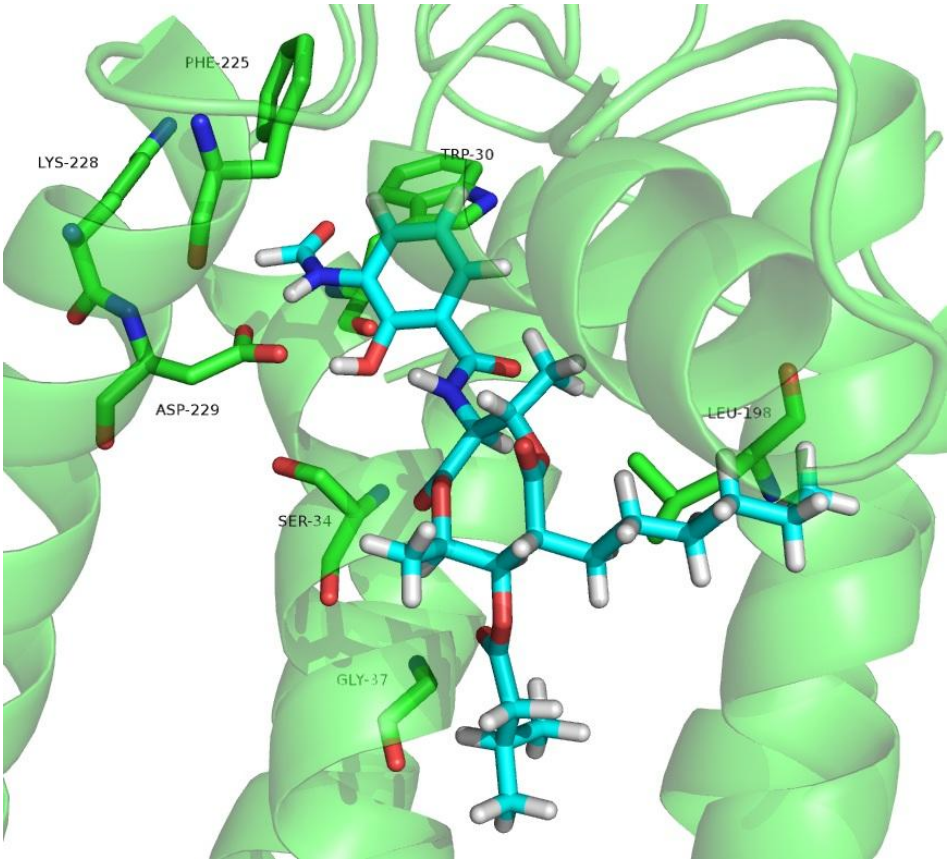


Fig. 5.23 Docking solution of antimycin A (shown in blue) in the Q_i pocket of the yeast cytochrome bc_1 complex (3CX5). Several hydrophobic and van der Waals interactions are observed. The yeast cytochrome b polypeptide backbone is represented in green.

Earlier it was discussed that two groups of residues have shown importance in Q_i binding. Figure 5.24 represents a mesh view of the protein bound with antimycin A. The Q_i binding site opens up into a large hydrophobic pocket, into which the side chains of antimycin A can move. Meanwhile, the amide substituted benzene ring can sit in a small pocket, opposite which resides another pocket, and it is believed that this alternative pocket preferentially binds alternative compounds.

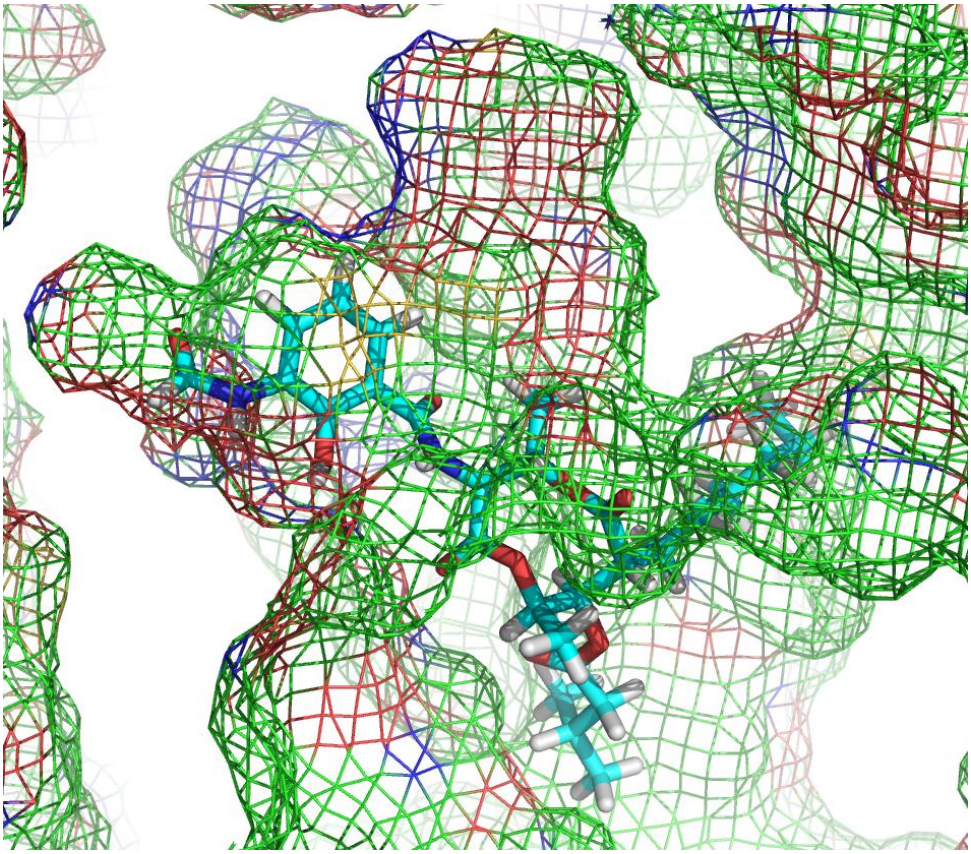


Fig. 5.24 Docking solution antimycin A (shown in blue) in the Q_i pocket of the yeast cytochrome bc_1 complex (3CX5). Protein represented as a mesh surface.

5.2.3.2 Docking of NQNO

Given that resistance can quickly hamper the clinical applications of a drug candidate, it is becoming increasingly more desirable to develop compounds that observe dual inhibition, that is, compounds which act at more than one site. Whilst SL-2-25 was found to inhibit two different enzymes in the ETC (*Pf*NDH2 and *Pf*bc_1), it is possible that a compound may act at two different sites in the same enzyme, such as the Q_o and the Q_i sites of the bc_1 complex. One such compound that has been found to observe this dual inhibition is 2-*n*-nonyl-4-hydroxyquinoline N-oxide (NQNO, fig. 5.25). NQNO contains a quinolone ring and is structurally similar to the natural substrate ubiquinone. Studies have shown that it is capable of

binding to both the Q_o and Q_i sites, signifying an important discovery in inhibitor binding studies of bc_1.[39, 40]

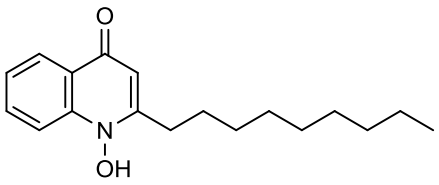


Fig. 5.25 NQNO.

To rationalise this dual binding a docking study was undertaken. NQNO was first energy minimised using the *'Energy Minimisation Protocol'*, and then docked into the Q_o and Q_i sites using the *'Q_o Docking Protocol'* and *'Q_i Docking Protocol'* respectively. Given the long alkyl side chain, the number of GA runs was increased from 10 to 25 to incorporate additional diversity in the solutions. Also, constraints were applied to bias the docking solutions at Q_o to form the His181 and Glu272 interactions.

The 25 docking poses at the Q_o site were all in strong agreement with one another, showing the quinolone ring residing tightly in the Q_o region, and the alkyl side chain moving into the hydrophobic pocket. The solutions had an average GOLDScore of 60.6 ± 2.4, and ChemScore of 35.1 ± 1.8. One solution is shown in figure 5.26 (GOLDScore and ChemScore values of 62.6 and 37.0 respectively). NQNO binds similar to SMA, with the quinolone head group inserted between residues Ile147 and Pro271. A strong H-bond is observed between the NH group of the His181 imidazole, and the carbonyl of the quinolone. The water mediated H-bond network is also observed between the OH group of NQNO and Glu272. The alky side chain also moves out into the hydrophobic pocket and forms a number of van der Waals contacts.

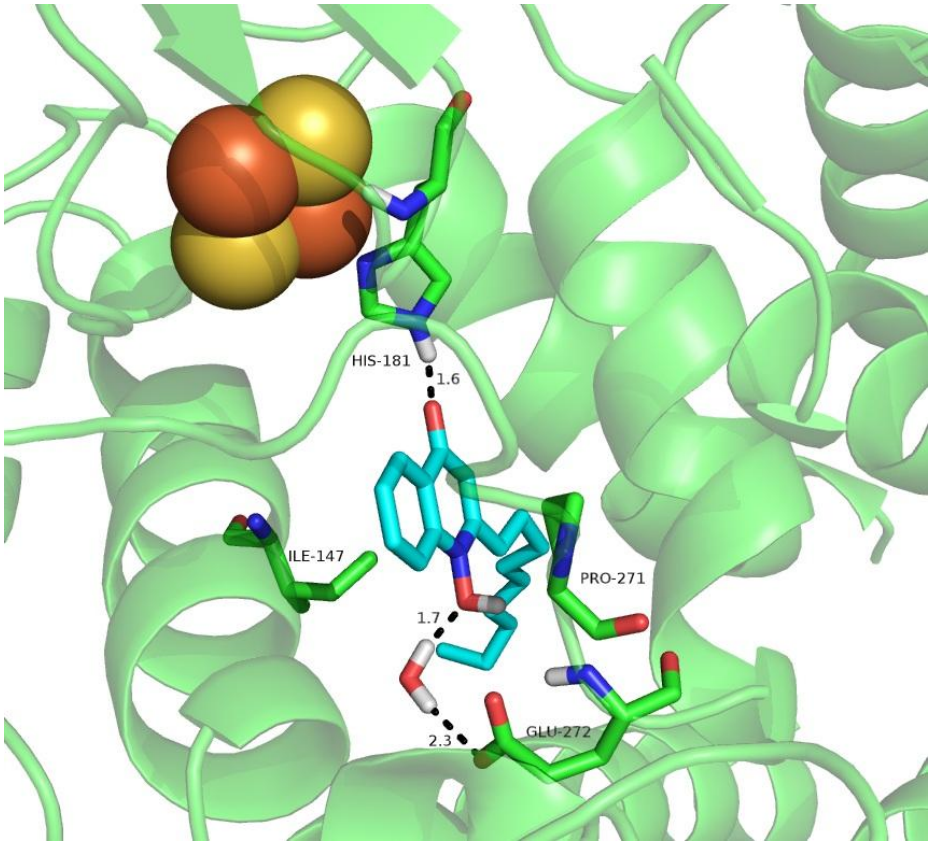


Fig. 5.26 Docking solution of NQNO (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 and quinolone carbonyl H-bond are clearly observed, as is the water mediated Glu272 interaction with the OH group of NQNO. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

The 25 docking solutions at the Q_i site were also all in strong agreement with one another, and were found to bind similar to antimycin A. The solutions had an average GOLDScore of 50.4 ± 1.7, and ChemScore of 27.6 ± 1.2. One solution is shown in figure 5.27 (GOLDScore and ChemScore values of 53.2 and 27.5 respectively). As with antimycin A, the hydroxyquinoline head group rests in a pocket comprised of residues Trp30, Phe225, Lys228 and Asp229. Interactions are also observed between Ser34 and Leu198, though the interaction with Gly37 is no longer seen as NQNO is too far away. Additionally, the alkyl chain appears to forms contacts with Tyr16, Ile195 and Met199, none of which were observed in antimycin A binding. This therefore supports the understanding that owing to the size and

shape of the Q_i pocket, inhibitors can bind in different ways, utilising the different subsets of residues.[39]

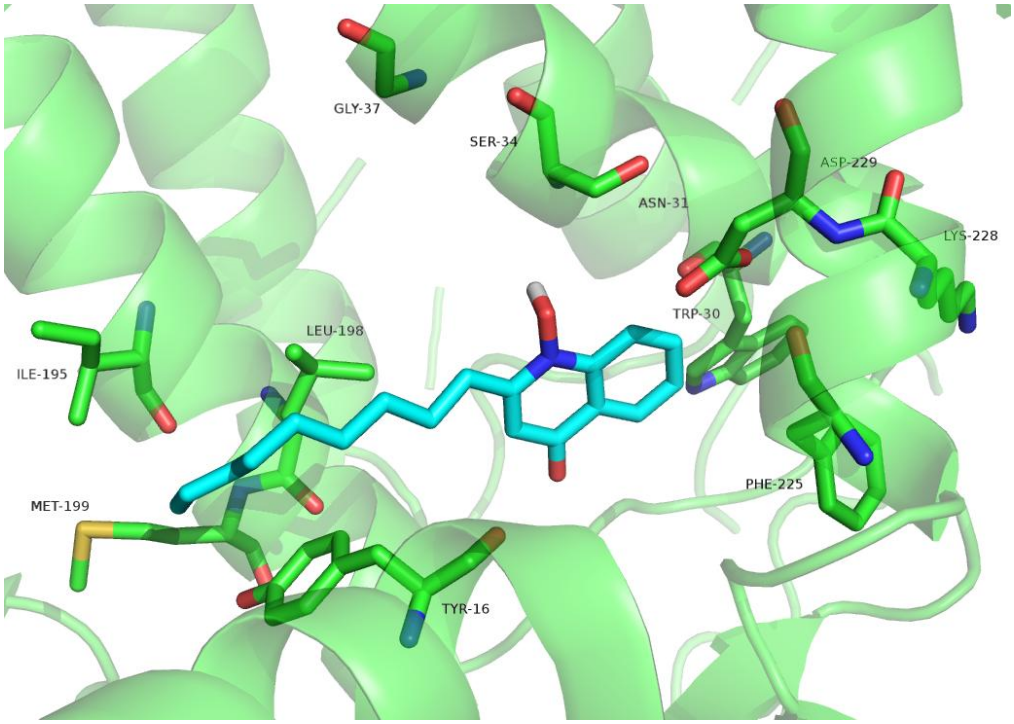


Fig. 5.27 Docking solution of NQNO (shown in blue) in the Q_i pocket of the yeast cytochrome bc_1 complex (3CX5). Several hydrophobic and van der Waals interactions are observed. The yeast cytochrome b polypeptide backbone is represented in green.

The docking study showed that NQNO could indeed bind to both the Q_o and Q_i sites, with visual inspection suggesting favourable orientations of the solutions, to provide the necessary H-bond, hydrophobic and van der Waals interactions to support activity. However, it would be highly beneficial if there were a quantitative means for assessing a compounds inhibition potential. The fitness scores may therefore provide additional support, as well as useful insight into the affinity of a compound at a particular target. The docking results from the NQNO study provided an ideal opportunity for such an analysis, with table 5.7 showing the average GOLDScore and ChemScore values.

Table. 5.7 Average GOLDScore and ChemScore values from the docking study of NQNO at the Q_o and Q_i sites of yeast bc_1.

Docking Site	Average GOLDScore Value	Average ChemScore Value
Q_o	60.6 ± 2.4	35.1 ± 1.8
Q_i	50.4 ± 1.7	27.6 ± 1.2

Though GOLDScore and ChemScore are unrelated to each other, independently they may be used to compare the docking scores of either different molecules at the same site, or the same molecule at different sites. This will allow for comment to be drawn as to the relative strength of binding of a particular molecule at a particular site. The average GOLDScore values for NQNO at the Q_o and Q_i sites were fairly similar to one another (60.6 and 50.4 respectively), suggesting that NQNO has broadly similar binding strengths at both sites. GOLDScore alone therefore suggests it has dual inhibition, which it does.[39, 40] The average GOLDScore at the Q_o site is slightly higher than that at Q_i however, but this is most likely due to the contribution of the strong H-bonds at Q_o, which strengthen the scores more so than weaker hydrophobic and van der Waals interactions. The same trend is also observed across the average ChemScore values.

Here molecular docking was used to support the understanding that NQNO is both a Q_o and Q_i binder, and through consideration of the fitness scores it was possible to quantify this observation. This line of investigation was subsequently expanded to see whether scoring functions could be used to predict whether a known bc_1 inhibitor was either class I or class II, without the need for extensive *in vitro* study. Inspection of the poses and their interactions could also potentially allow for conclusions to be drawn as to their binding modes.

5.2.3.3 Docking of HDQ

To test this hypothesis a docking study was performed on HDQ (fig. 5.28). HDQ is identical to NQNO, albeit with the alkyl side chain extended by two carbons. It is shown to inhibit *P. falciparum* replication in nM concentrations,[90] and given its structural similarity with ubiquinol, seems likely to target ubiquinol binding sites of respiratory enzymes. This is indeed the case as it is a highly potent *Pf*NDH2 inhibitor,[91] though its antimalarial use is limited owing to its poor drug like properties.[90, 92] Previous work has also suggested that HDQ may inhibit other ubiquinone dependent pathways in addition to *Pf*NDH2.[92]

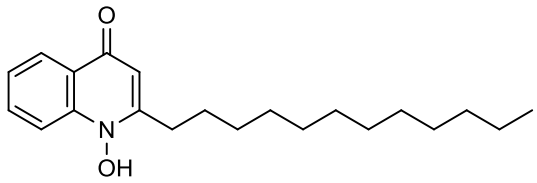


Fig. 5.28 HDQ.

The HDQ related compound 2-heptyl-4-hydroxyquinoline N-oxide (HQNO, fig. 5.29) is a known inhibitor of the mammalian and *S. cerevisiae* bc_1 complex. Several mutations causing resistance to HQNO have been reported in yeast, all of which are located in the Q_i region of the bc_1 complex.[93] From this observation questions were raised as to whether the antimalarial activity of HDQ was due to inhibition of the bc_1 complex, more specifically the Q_i site. Single point mutations of the Q_o site have been responsible for ATOV resistance. However, it has been shown that for the parasite bc_1 complex, both control and ATOV resistant strains are inhibited in sub μM concentrations by HDQ. This suggests that ATOV and HDQ have different targets within the bc_1 complex. Further to this, the introduction of point mutations in

the Q_i site in yeast (G33A, H204Y, M221Q and K228M) led to a marked decrease in HDQ inhibition.

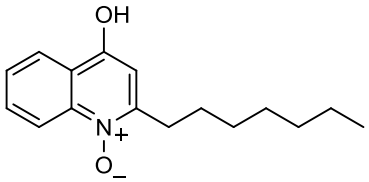


Fig. 5.29 HQNO.

With these biological observations in hand, molecular docking was employed in an attempt to rationalise and offer support for these results. HDQ was drawn and energy minimised using the *'Energy Minimisation Protocol'*, and then docked into both the Q_o and Q_i binding sites using the *'Q_o Docking Protocol'* and *'Q_i Docking Protocol'* respectively. To complement the study, the same docking was simultaneously performed for SMA, allowing for a comparison of the fitness scores. The average GOLDScore and ChemScore values for SMA and HDQ across the 10 GA solutions at both the Q_o and Q_i sites are shown in table 5.8.

Table. 5.8 HDQ and SMA docking results at both the Q_o and Q_i sites.

	Q_o		**Q_i**	
	SMA	HDQ	SMA	HDQ
Average GOLDScore	92.3 ± 13.9	63.5 ± 1.3	45.4 ± 13.0	54.1 ± 3.5
Average ChemScore	54.0 ± 3.4	37.7 ± 1.2	28.0 ± 4.7	28.6 ± 2.1

Inspection of table 5.8 shows some interesting observations with regard to the fitness scores. First off, the GOLDScore and ChemScore values were both significantly higher for SMA at the Q_o site, as oppose to Q_i. Based on the scores alone this would suggest that SMA binds preferentially to the Q_o site, which is indeed known to be the case.[39] Though this is encouraging, the additional H-bonds at the Q_o site contribute greatly to the fitness scores, therefore the higher values were expected. Also, the Q_o site is perfectly shaped for SMA as it is the native ligand. It therefore may be more

meaningful to comment on the docking scores of different molecules at the same sites. To this ends, it can be seen from table 5.8 that both the average GOLDScore and ChemScore values at the Q_o site are considerable lower for HDQ, than SMA. From this it would be reasonable to say that SMA binds more strongly at Q_o than HDQ. Converse to this, the average GOLDScore and ChemScore values at Q_i would suggest that HDQ is a stronger binder at Q_i than SMA. This is particularly evident with GOLDScore. Using this docking study in combination with the knowledge that SMA is a known, potent Q_o inhibitor would suggest that HDQ does indeed exert a large proportion of its bc_1 inhibition effect through binding at Q_i.

Further validation of this could be garnered by consideration of the binding modes for HDQ. Figure 5.30 represents one solution of HDQ docked into the Q_o site. As was expected, the quinolone head group resides nicely in the Q_o site, with the alkyl side chain moving out into the hydrophobic pocket. A strong H-bond between the carbonyl of the quinolone and the NH of the His181 imidazole is observed, as is the potential for a water mediated Glu272 interaction with the OH of HDQ. Yet in comparison to SMA, the interactions are clearly not as strong and significant.

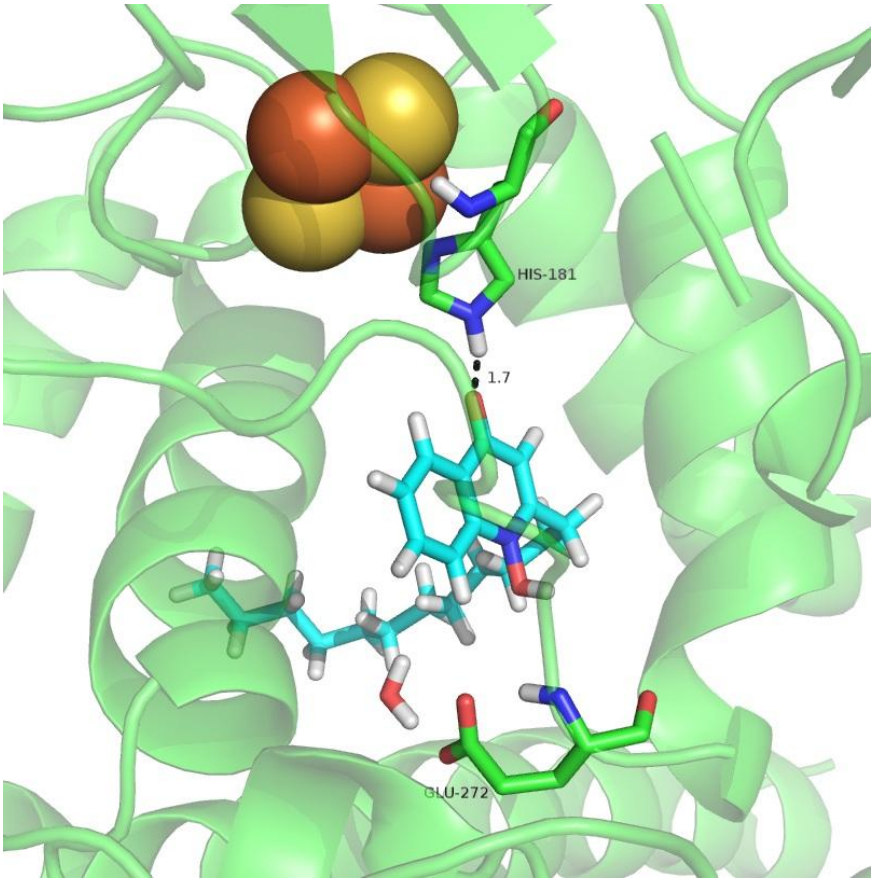


Fig. 5.30 Docking solution of HDQ (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with the quinolone carbonyl is clearly observed, as is the potential for a water mediated interaction with Glu272. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Of more importance is the binding mode of HDQ in the Q_i site. Figure 5.31 represents one such docking solution, though all of the poses bound similarly. The hydroxyl moiety of Ser34 was shown to potentially form a H-bond to the carbonyl group of HDQ, along with another H-bond between the HDQ N-oxide, and the hydroxyl moiety of residue Ser206. Additionally, a H-bond was also predicted to form between Asp229 and bound HDQ. Several weaker van der Waals and hydrophobic interactions were also observed, such as Tyr16 and Met221.

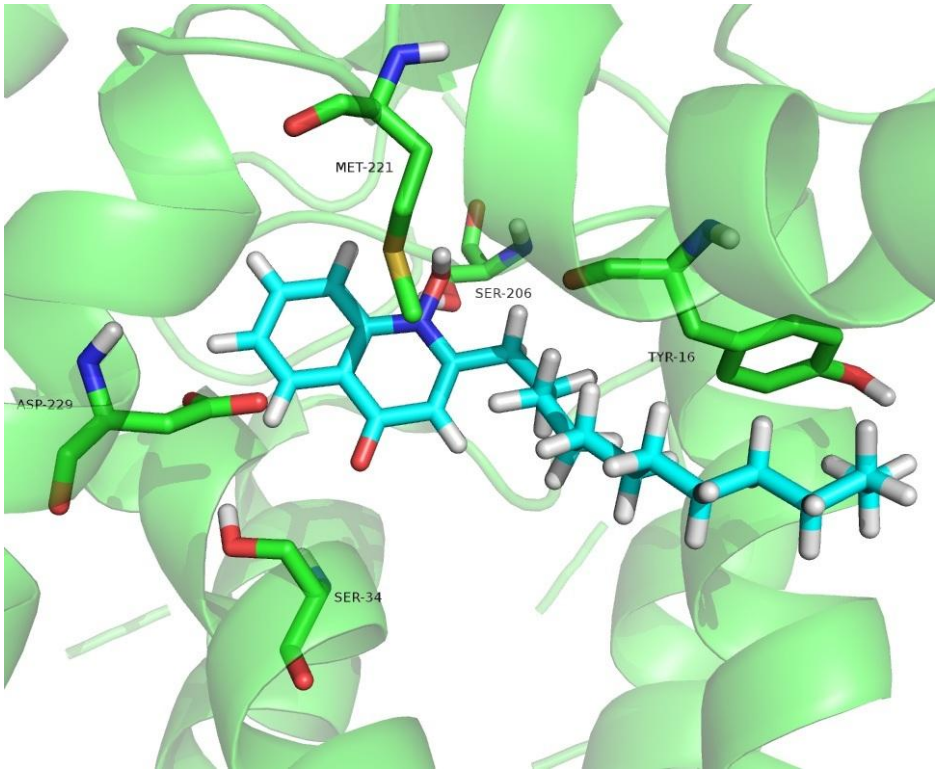


Fig. 5.31 Docking solution of HDQ (shown in blue) in the Q_i pocket of the yeast cytochrome bc_1 complex (3CX5). Several potential H-bond, hydrophobic and van der Waals interactions are shown. The yeast cytochrome b polypeptide backbone is represented in green.

Analysis of the fitness scores and docking solutions seem to suggest that HDQ is indeed a Q_i binder. To offer further support, the native Q_i ligand ubiquinone (fig. 5.32) was similarly docked into the Q_i site using the *'Q_i Docking Protocol'*.

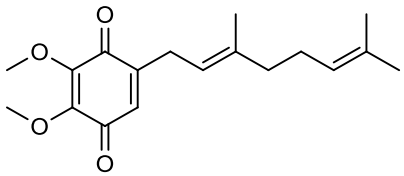


Fig. 5.32 Ubiquinone.

The results of this docking are shown in table 5.9, along with the fitness scores of SMA and HDQ for comparison. Given that ubiquinone is the native ligand of the Q_i site, its fitness scores were used as a point of reference from which comparisons could be made. Ubiquinone had an average GOLDScore of 48.3 ± 1.4 across the 10 GA runs, and was therefore comparable to that of SMA (45.4 ± 13.0). HDQ on the

other hand clearly bound more strongly than both SMA and ubiquinone, with an average GOLDScore of 54.1 ± 3.5. This tight binding could potentially explain the observed cardiotoxicity of HDQ in mammals. Unfortunately however, ChemScore failed to replicate this trend, though this is not too discouraging as it should be routine to investigate multiple scoring functions for different applications.[16, 17]

Table. 5.9 Ubiquinone, HDQ and SMA docking results at the Q_i site.

	Q_i		
	SMA	HDQ	Ubiquinone
Average GOLDScore	45.4 ± 13.0	54.1 ± 3.5	48.3 ± 1.4
Average ChemScore	28.0 ± 4.7	28.6 ± 2.1	22.3 ± 3.8

By combining insight garnered from the study of yeast with specific cytochrome b mutations, and by using molecular docking, it has been shown that HDQ inhibition at the cytochrome bc_1 complex is most significantly mediated via Q_i binding. Therefore, HDQ displays a novel inhibitory mode of action and offers further support that the Q_i site of *Pf*bc_1 is a viable target for antimalarial drug development. This docking study formed part of a publication in the journal of *'Antimicrobial Agents and Chemotherapy'* entitled *'HDQ, a potent inhibitor of Plasmodium falciparum proliferation, binds to the quinone reduction site of the cytochrome bc_1 complex'*.[94]

5.2.4 Docking of LBVS Hits

To supplement the detailed analysis of the LBVS hits at the end of Chapter IV, the five compounds (VS01, VS09, VS10, VS16, VS18) which were found to be active against the malaria parasite (table 4.5) were docked into the Q_o and Q_i sites of the yeast bc_1 complex. This was done to gather further information in support of the hypothesis that they are *Pf*bc_1 inhibitors, given that the biological testing which

would definitively prove this was unavailable. The five compounds were constructed and energy minimised using the *'Energy Minimisation Protocol'*, and then docked into the Q_o and Q_i sites using the *'Q_o Docking Protocol'* and *'Q_i Docking Protocol'* respectively. The average docking scores for each of the compounds at Q_o and Q_i are shown in tables 5.10 and 5.11 respectively.

Table. 5.10 Fitness function scores for the docking results of the LBVS hits at the Q_o site.

	VS01	**VS09**	**VS10**	**VS16**	**VS18**
Average GOLDScore	47.1 ± 6.2	53.3 ± 0.8	52.6 ± 3.5	64.9 ± 2.2	57.9 ± 4.4
Average ChemScore	30.0 ± 2.2	20.2 ± 0.5	19.8 ± 6.0	29.1 ± 1.7	30.7 ± 1.8

Table. 5.11 Fitness function scores for the docking results of the LBVS hits at the Q_i site.

	VS01	**VS09**	**VS10**	**VS16**	**VS18**
Average GOLDScore	49.7 ± 3.6	48.6 ± 0.5	55.2 ± 1.5	48.1 ± 2.6	47.3 ± 0.6
Average ChemScore	26.6 ± 4.1	17.2 ± 1.3	27.5 ± 0.7	20.8 ± 1.7	23.9 ± 1.2

The scores were quite similar across the five compounds; therefore the solutions were visually inspected in order to draw comment as to the nature of their binding. At Q_o, VS01 (fig. 4.4) was found to bind according to figure 5.33. A weak H-bond was observed between the NH group of His181 and a carbonyl group on the pyrrolopyrimidine-2,4-dione core. However, owing to the bulk of the phenyl groups and the narrow entrance of the Q_o site, the compound could not orientate itself tightly in the pocket, thus the water mediate interaction with Glu272 was not seen. The presence of the His181 interaction clearly supports the assumption that this compound is a *Pf*bc_1 inhibitor, but its otherwise poor binding may explain its lack of reproducibility *in vitro*. Though Q_i docking was less informative, there was the potential for H-bonding between a carbonyl group in VS01, and the hydroxyl moiety of Ser206 and the amine group of Gln22, similar to the observed binding of HDQ (fig. 5.31).

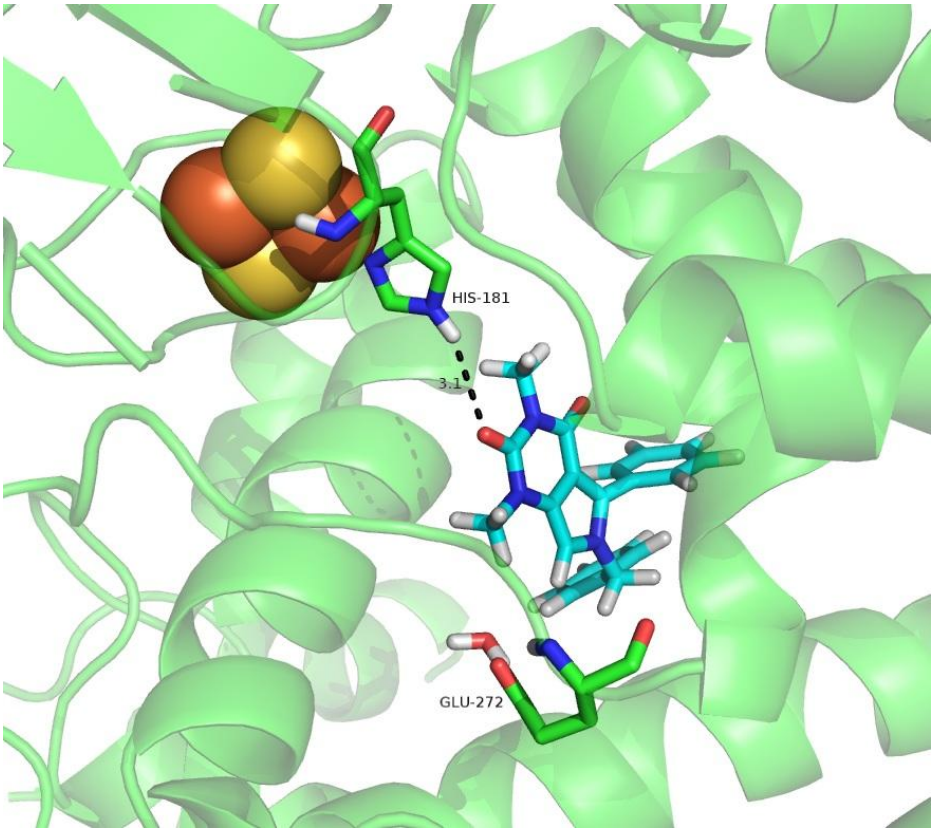


Fig. 5.33 Docking solution of VS01 (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with one of the pyrrolopyrimidine-2,4-dione carbonyl groups is clearly shown. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Of the five compounds, VS09 (fig. 4.6) appeared to be the most promising from docking analysis. A representative pose for VS09 is shown in figure 5.34. There was a strong H-bond between His181 and the nitrogen of the benzothiazole ring, as well as a water mediated H-bonding network between the imidazole side chain and Glu272. The strength of these bonds may go some way to rationalising the relatively potent 3D7 inhibition of this compound (IC_{50} = 4.53 ± 1.86 µM), as well as its good *in vitro* reproducibility. Given the size of this compound and its positioning in the Q_o active site, there is clearly much scope for chemical optimisation with regard to this chemotype, and all things considered it looks to be the most promising lead like structure from the virtual screening study.

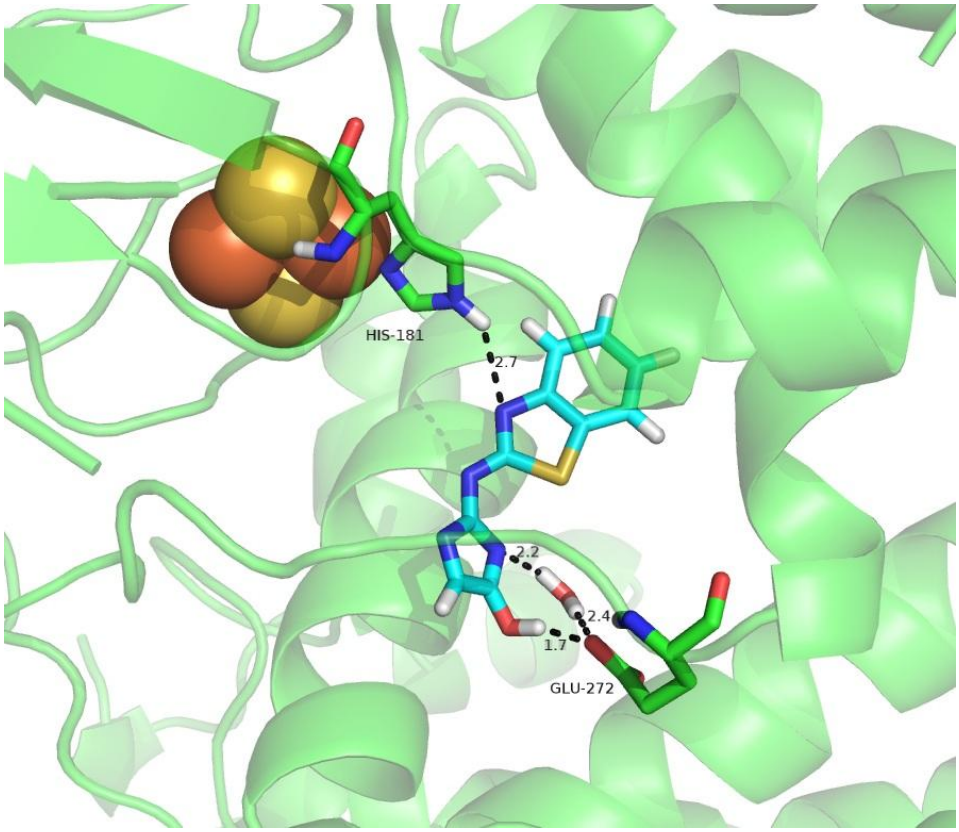


Fig. 5.34 Docking solution of VS09 (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with the benzothiazole nitrogen is clearly shown, as well as the potential for a water mediated interaction with Glu272 and the side chain of VS09. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

VS09 also docked very well into the Q_i site, suggesting that it may have dual inhibition at both sites, similar to NQNO. A strong H-bond was observed between the hydroxyl group of its side chain and the Asp229 amino acid, as well as a number of potential H-bonds between the Ser206 and Gln22 protein residues and the nitrogen of the benzothiazole chemotype and nitrogen linker, as observed in figure 5.35.

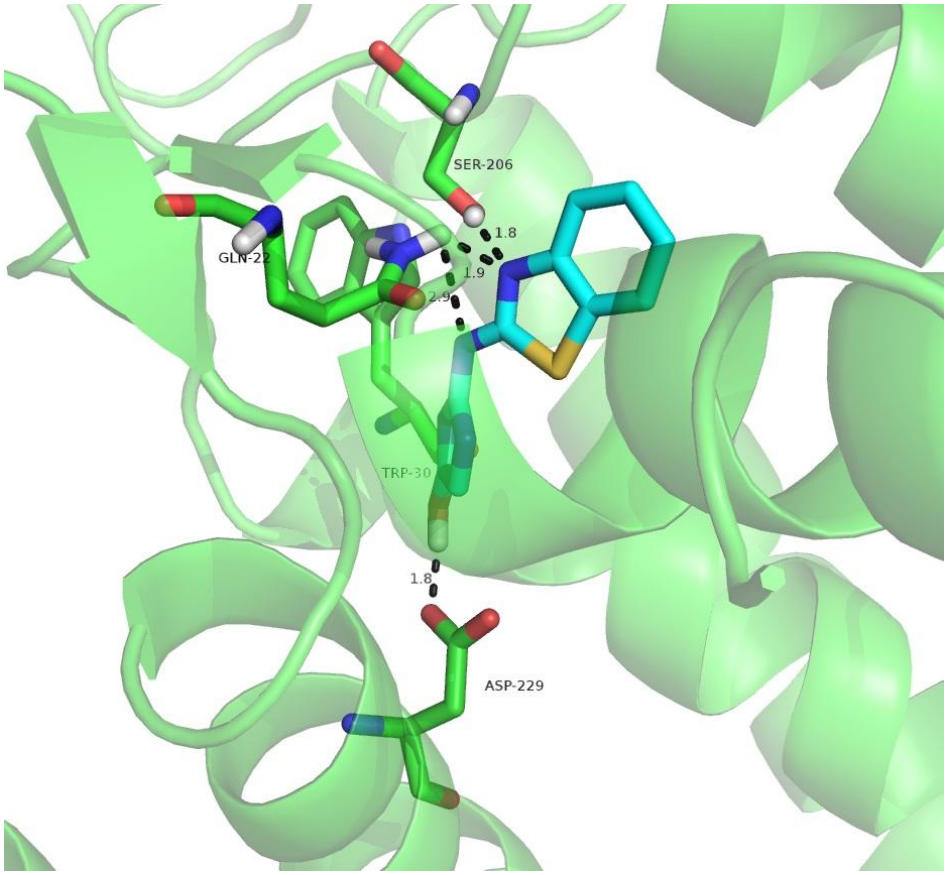


Fig. 5.35 Docking solution of VS09 (shown in blue) in the Q_i pocket of the yeast cytochrome bc_1 complex (3CX5). Several hydrophobic and van der Waals interactions are observed, as well as potential H-bond contacts. The yeast cytochrome b polypeptide backbone is represented in green. H-bonds are indicated by black lines.

The rigid isoalloxazine ring of VS10 (fig. 4.11) docked into the Q_o site, forming a H-bond between one of its carbonyl groups and His181, as shown in figure 5.36. The strength of this interaction helped to rationalise its observed activity (IC_{50} = 6.41 ± 1.73 µM), and the potential of the isoalloxazine chemotype as a lead like structure. Similar to VS01 and VS09 (fig. 5.35), the carbonyl group of VS10 was able to H-bond to Ser206 and Gln22 at the Q_i site.

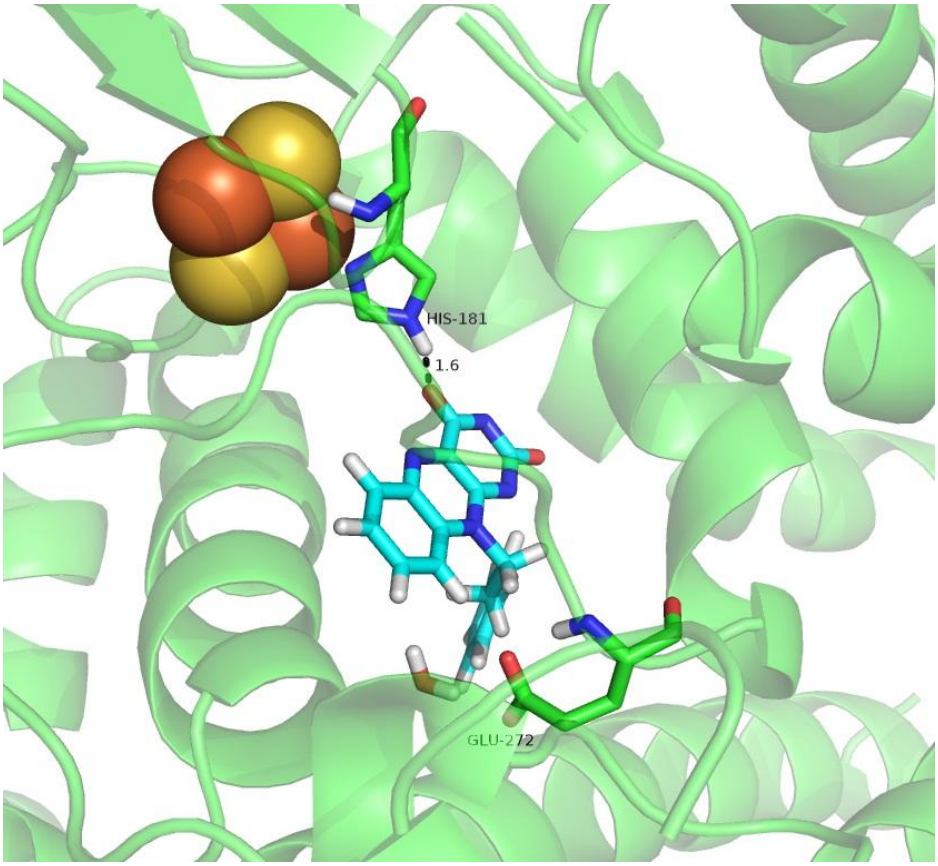


Fig. 5.36 Docking solution of VS10 (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with the isoalloxazine ring is clearly shown. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

Neither VS16 nor VS18 (fig. 4.15) contain traditional chemotypes (i.e. heterocyclic structures with unique functionality), thus their docking at the Q_o site was fairly unremarkable. However, both were found to observe strong H-bond interactions between a carbonyl group and His181, as shown in figures 5.37 and 5.38 respectively. These observations may go some way to explaining their *in vitro* activities. Unfortunately, when docked at the Q_i site, though their docking scores were comparable to the other hits, no significant interactions were observed.

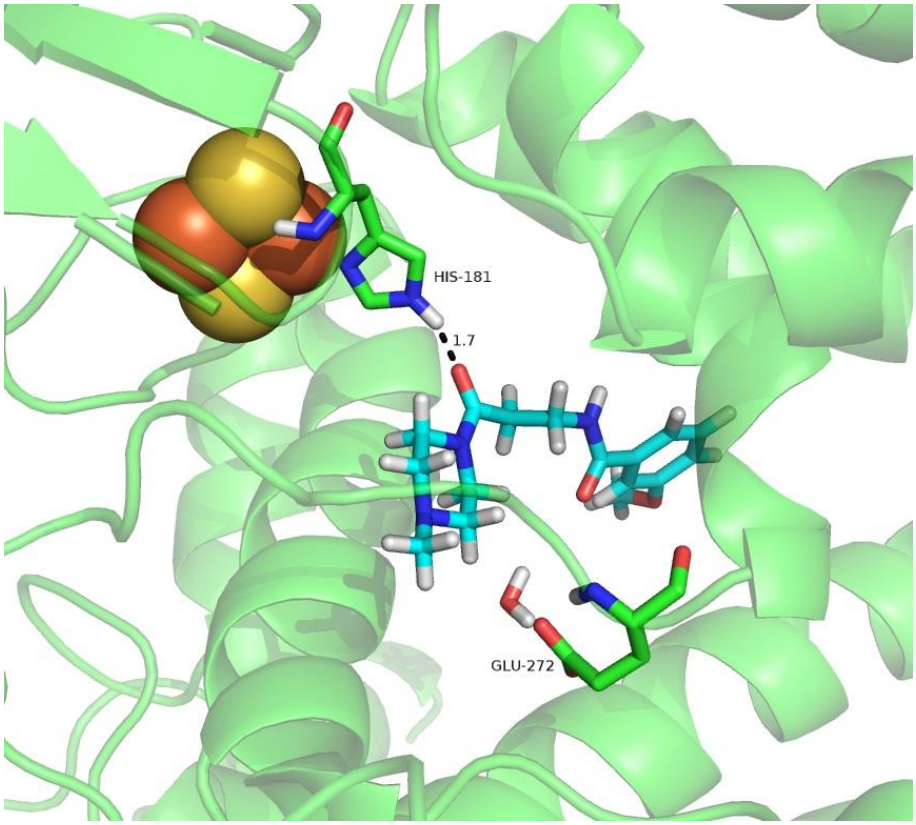


Fig. 5.37 Docking solution of VS16 (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with a carbonyl group is clearly shown. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

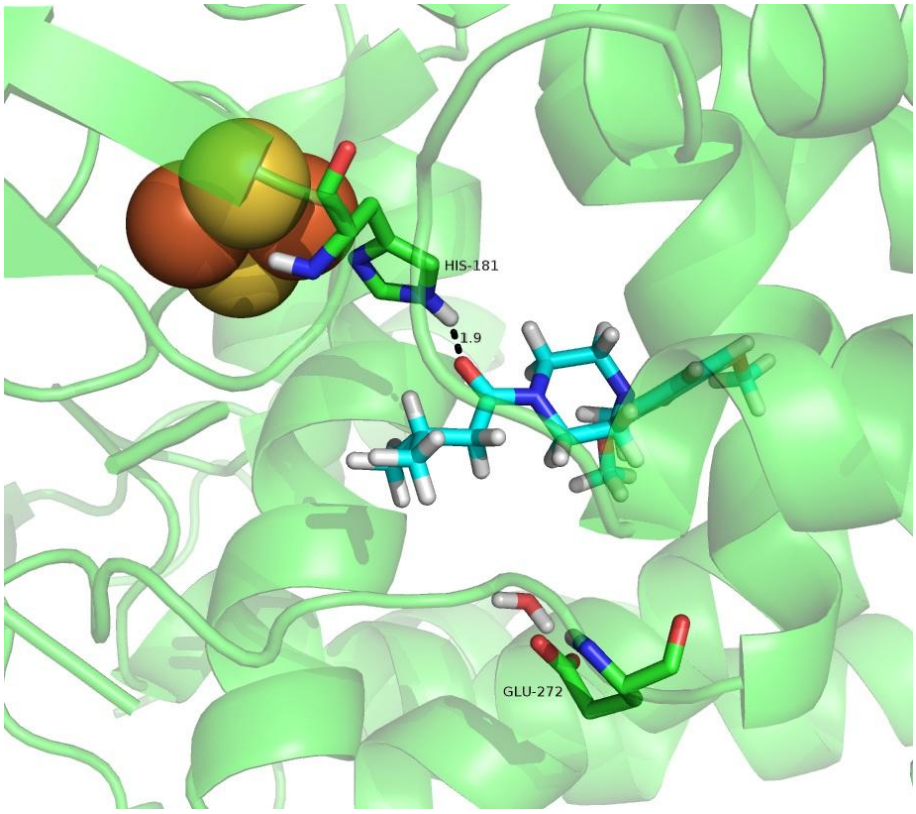


Fig. 5.38 Docking solution of VS18 (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with a carbonyl group is clearly shown. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

By combining this docking analysis with the earlier interpretation of the chemotypes and biological results (Chapter IV), it is indeed most likely that these compounds are inhibiting the *Pf*bc_1 complex. The docking therefore offers further support of the strength of the LBVS work which was performed, and by extension, that a number of exciting lead like chemotypes have been identified.

5.3 Summary of Docking Studies

Combined, these studies show that molecular docking has many applications in modern drug discovery. Not only can it be used to explain a compounds activity profile through consideration of observed interactions, but it has also proven useful in confirming the identity of novel antimalarial targets, and in rationalising

observations with regard to emerging resistance. Clearly there is much potential and scope to utilise molecular docking tools to aid in the systematic design of new antimalarial compounds, active against the cytochrome bc_1 complex. With work ongoing to design and validate a homology model of *Pf*bc_1, docking will help to drive forward ongoing efforts to further refine the molecular design loop.

Chapter VI will discuss the application of QSAR methods to tackle two problems. Firstly, to develop QSARs that could be used to predict the activity of 4-aminoquinoline compounds against CQ sensitive and CQ resistant strains of malaria. And secondly, to develop models for the assessment of drug safety for a series of thiazolide compounds active against hepatitis C. QSAR methods will be fully introduced and discussed, together with appropriate validation techniques.

5.4 References

1. M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins*, 2003, **52**, 609-623.
2. T. Lengauer and M. Rarey, *Curr. Opin. Struct. Biol.*, 1996, **6**, 402-406.
3. D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat Rev Drug Discov*, 2004, **3**, 935-949.
4. J. M. Blaney and J. S. Dixon, *Perspect. Drug Discov. Design*, 1993, **1**, 301-319.
5. R. Abagyan and M. Totrov, *Curr. Opin. Chem. Biol.*, 2001, **5**, 375-382.
6. R. D. Taylor, P. J. Jewsbury and J. W. Essex, *Journal of Computer-Aided Molecular Design*, 2002, **16**, 151-166.
7. I. Halperin, B. Y. Ma, H. Wolfson and R. Nussinov, *Proteins*, 2002, **47**, 409-443.
8. S. Makino and I. D. Kuntz, *Journal of Computational Chemistry*, 1997, **18**, 1812-1825.
9. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *Journal of Molecular Biology*, 1996, **261**, 470-489.
10. C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead and M. D. Eldridge, *Proteins*, 1998, **33**, 367-382.
11. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *Journal of Molecular Biology*, 1997, **267**, 727-748.
12. G. Jones, P. Willett and R. C. Glen, *Journal of Molecular Biology*, 1995, **245**, 43-53.
13. *GOLD 5.0.1*, CCDC Software Limited 2005-2010, http://www.ccdc.cam.ac.uk/products/.
14. J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins*, 2002, **49**, 457-471.
15. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235-242.
16. M. Kontoyianni, L. M. McClellan and G. S. Sokol, *Journal of Medicinal Chemistry*, 2004, **47**, 558-565.
17. G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, *Journal of Medicinal Chemistry*, 2006, **49**, 5912-5931.
18. Ajay and M. A. Murcko, *Journal of Medicinal Chemistry*, 1995, **38**, 4953-4967.
19. R. Rajamani and A. C. Good, *Current Opinion in Drug Discovery and Development*, 2007, **10**, 308-315.
20. M. H. J. Seifert, J. Kraus and B. Kramer, *Current Opinion in Drug Discovery and Development*, 2007, **10**, 298-307.
21. A. N. Jain, *Current Protein and Peptide Science*, 2006, **7**, 407-420.
22. O. Korb, T. Stützle and T. E. Exner, *Journal of Chemical Information and Modeling*, 2009, **49**, 84-96.
23. I. Muegge and Y. C. Martin, *Journal of Medicinal Chemistry*, 1999, **42**, 791-804.
24. J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster and J. M. Thornton, *Journal of Computational Chemistry*, 1999, **20**, 1177-1185.
25. H. Gohlke, M. Hendlich and G. Klebe, *Journal of Molecular Biology*, 2000, **295**, 337-356.
26. C. A. Sotriffer, H. Gohlke and G. Klebe, *Journal of Medicinal Chemistry*, 2002, **45**, 1967-1970.
27. I. Muegge, *Journal of Medicinal Chemistry*, 2005, **49**, 5895-5902.
28. P. K. Weiner and P. A. Kollman, *Journal of Computational Chemistry*, 1981, **2**, 287-303.
29. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 425-445.
30. C. W. Murray, T. R. Auton and M. D. Eldridge, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 503-519.
31. H. J. Böhm, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 309-323.
32. V. Barton, N. Fisher, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Curr. Opin. Chem. Biol.*, 2010, **14**, 440-446.
33. M. Fry and M. Pudney, *Biochem. Pharmacol.*, 1992, **43**, 1545-1553.
34. P. Mitchell, *FEBS Lett.*, 1975, **59**, 137-139.
35. A. R. Crofts, *Annu. Rev. Physiol.*, 2004, **66**, 689-733.
36. I. K. Srivastava, H. Rottenberg and A. B. Vaidya, *J. Biol. Chem.*, 1997, **272**, 3961-3966.
37. G. A. Biagini, P. Viriyavejakul, P. M. O'Neill, P. G. Bray and S. A. Ward, *Antimicrob. Agents Chemother.*, 2006, **50**, 1841-1851.

38. J. L. Cape, M. K. Bowman and D. M. Kramer, *Trends Plant Sci.*, 2006, **11**, 46-55.
39. L. Esser, B. Quinn, Y. F. Li, M. Q. Zhang, M. Elberry, L. Yu, C. A. Yu and D. Xia, *Journal of Molecular Biology*, 2004, **341**, 281-302.
40. X. G. Gao, X. L. Wen, L. Esser, B. Quinn, L. Yu, C. A. Yu and D. Xia, *Biochemistry*, 2003, **42**, 9067-9080.
41. S. R. N. Solmaz and C. Hunte, *J. Biol. Chem.*, 2008, **283**, 17542-17549.
42. R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. Oneill, *MedChemComm*, 2012, **3**.
43. T. A. Link, M. Iwata, J. Bjorkman, D. van der Spoel, A. Stocker and S. Iwata, *Molecular modeling of inhibitors at Q(i) and Q(o) sites in cytochrome bc(1) complex*, Wiley-Vch, Inc, New York, 2003.
44. T. Ohnishi, U. Brandt and G. Vonjagow, *Eur. J. Biochem.*, 1988, **176**, 385-389.
45. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
46. M. Degliesposti, S. Devries, M. Crimi, A. Ghelli, T. Patarnello and A. Meyer, *Biochimica Et Biophysica Acta*, 1993, **1143**, 243-271.
47. C. Hunte, J. Koepke, C. Lange, T. Rossmanith and H. Michel, *Struct. Fold. Des.*, 2000, **8**, 669-684.
48. T. A. Link, W. R. Hagen, A. J. Pierik, C. Assmann and G. Vonjagow, *Eur. J. Biochem.*, 1992, **208**, 685-691.
49. A. R. Crofts, S. J. Hong, N. Ugulava, B. Barquera, R. Gennis, M. Guergova-Kuras and E. A. Berry, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 10021-10026.
50. T. A. Link, *FEBS Lett.*, 1997, **412**, 257-264.
51. S. Junemann, P. Heathcote and P. R. Rich, *J. Biol. Chem.*, 1998, **273**, 21603-21607.
52. C. H. Snyder, E. B. Gutierrez-Cirlos and B. L. Trumpower, *J. Biol. Chem.*, 2000, **275**, 13535-13541.
53. M. Brugna, S. Rodgers, A. Schricker, G. Montoya, M. Kazmeier, W. Nitschke and I. Sinning, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 2069-2074.
54. C. Lange and C. Hunte, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 2800-2805.
55. M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, J. W. M. Nissink, R. D. Taylor and R. Taylor, *Journal of Medicinal Chemistry*, 2005, **48**, 6504-6515.
56. *PyMOL*, http://www.pymol.org/.
57. R. S. Judson, E. P. Jaeger and A. M. Treasurywala, *Theochem-J. Mol. Struct.*, 1994, **114**, 191-206.
58. C. M. Oshiro, I. D. Kuntz and J. S. Dixon, *Journal of Computer-Aided Molecular Design*, 1995, **9**, 113-130.
59. D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel and S. T. Freer, *Chem. Biol.*, 1995, **2**, 317-324.
60. J. J. Kessl, S. R. Meshnick and B. L. Trumpower, *Trends in Parasitology*, 2007, **23**, 494-501.
61. M. W. Mather, E. Darrouzet, M. Valkova-Valchanova, J. W. Cooley, M. T. McIntosh, F. Daldal and A. B. Vaidya, *J. Biol. Chem.*, 2005, **280**, 27458-27465.
62. J. Krungkrai, S. R. Krungkrai, N. Suraveratum and P. Prapunwattana, *Biochem. Mol. Biol. Int.*, 1997, **42**, 1007-1014.
63. M. W. Mather and A. B. Vaidya, *J. Bioenerg. Biomembr.*, 2008, **40**, 425-433.
64. J. J. Kessl, B. B. Lange, T. Merbitz-Zahradnik, K. Zwicker, P. Hill, B. Meunier, H. Palsdottir, C. Hunte, S. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2003, **278**, 31312-31318.
65. J. J. Kessl, N. V. Moskalev, G. W. Gribble, M. Nasr, S. R. Meshnick and B. L. Trumpower, *Biochim. Biophys. Acta-Bioenerg.*, 2007, **1767**, 319-326.
66. Spartan, *Wavefunction, INC*, 2008.
67. T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490-519.
68. M. Korsinczky, N. H. Chen, B. Kotecka, A. Saul, K. Rieckmann and Q. Cheng, *Antimicrob. Agents Chemother.*, 2000, **44**, 2100-2108.
69. L. Musset, O. Bouchaud, S. Matheron, L. Massias and J. Le Bras, *Microbes Infect.*, 2006, **8**, 2599-2604.
70. A. Berry, A. Senescau, J. Lelievre, F. Benoit-Vical, R. Fabre, B. Marchou and J. F. Magnaval, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 2006, **100**, 986-988.
71. N. Fisher and B. Meunier, *FEMS Yeast Res.*, 2008, **8**, 183-192.

72. Q. L. Fivelman, G. A. Butcher, I. S. Adagu, D. C. Warhurst and G. Pasvol, *Malaria Journal*, 2002, **1**, 1.
73. E. Schwartz, S. Bujanover and K. C. Kain, *Clin. Infect. Dis.*, 2003, **37**, 450-451.
74. N. Fisher and B. Meunier, *Pest Manag. Sci.*, 2005, **61**, 973-978.
75. J. J. Kessl, K. H. Ha, A. K. Merritt, B. B. Lange, P. Hill, B. Meunier, S. R. Meshnick and B. L. Trumpower, *J. Biol. Chem.*, 2005, **280**, 17142-17148.
76. N. Fisher, R. Abd Majid, T. Antoine, M. Al-Helal, A. J. Warman, D. J. Johnson, A. S. Lawrenson, H. Ranson, P. M. O'Neill, S. A. Ward and G. A. Biagini, *The Journal of biological chemistry*, 2012, **287**, 9731-9741.
77. A. R. Crofts, M. Guergova-Kuras, R. Kuras, N. Ugulava, J. Y. Li and S. J. Hong, *Biochim. Biophys. Acta-Bioenerg.*, 2000, **1459**, 456-466.
78. E. A. Berry and L.-S. Huang, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2011, **1807**, 1349-1363.
79. N. Fisher, C. K. Castleden, I. Bourges, G. Brasseur, G. Dujardin and B. Meunier, *J. Biol. Chem.*, 2004, **279**, 12951-12958.
80. F. Wibrand, K. Ravn, M. Schwartz, T. Rosenberg, N. Horn and J. Vissing, *Ann. Neurol.*, 2001, **50**, 540-543.
81. R. W. Winter, J. X. Kelly, M. J. Smilkstein, R. Dodean, D. Hinrichs and M. K. Riscoe, *Exp. Parasitol.*, 2008, **118**, 487-497.
82. A. F. G. Slater and A. Cerami, *Nature*, 1992, **355**, 167-169.
83. R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. Oneill, *MedChemComm*, 2012, **3**, 39-44.
84. E. Stern, G. G. Muccioli, R. Millet, J. F. Goossens, A. Farce, P. Chavatte, J. H. Poupaert, D. M. Lambert, P. Depreux and J. P. Henichart, *Journal of Medicinal Chemistry*, 2006, **49**, 70-79.
85. C. G. Wang, T. Langer, P. G. Kamath, Z. Q. Gu, P. Skolnick and R. I. Fryer, *Journal of Medicinal Chemistry*, 1995, **38**, 950-957.
86. T. Furuta, T. Sakai, T. Senga, T. Osawa, K. Kubo, T. Shimizu, R. Suzuki, T. Yoshino, M. Endo and A. Miwa, *Journal of Medicinal Chemistry*, 2006, **49**, 2186-2192.
87. R. Peter R, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2004, **1658**, 165-171.
88. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
89. G. A. Biagini, N. Fisher, A. E. Shone, M. A. Mubaraki, A. Srivastava, A. Hill, T. Antoine, A. J. Warman, J. Davies, C. Pidathala, R. K. Amewu, S. C. Leung, R. Sharma, P. Gibbons, D. W. Hong, B. Pacorel, A. S. Lawrenson, S. Charoensutthivarakul, L. Taylor, O. Berger, A. Mbekeani, P. A. Stocks, G. L. Nixon, J. Chadwick, J. Hemingway, M. J. Delves, R. E. Sinden, A.-M. Zeeman, C. H. M. Kocken, N. G. Berry, P. M. O'Neill and S. A. Ward, *Proceedings of the National Academy of Sciences*, 2012, **109**, 8298-8303.
90. A. Saleh, J. Friesen, S. Baumeister, U. Gross and W. Bohne, *Antimicrob. Agents Chemother.*, 2007, **51**, 1217-1222.
91. A. Eschemann, A. Galkin, W. Oettmeier, U. Brandt and S. Kerscher, *J. Biol. Chem.*, 2005, **280**, 3138-3142.
92. S. S. Lin, S. Kerscher, A. Saleh, U. Brandt, U. Gross and W. Bohne, *Biochim. Biophys. Acta-Bioenerg.*, 2008, **1777**, 1455-1462.
93. G. Brasseur, A. S. Saribas and F. Daldal, *Biochim. Biophys. Acta-Bioenerg.*, 1996, **1275**, 61-69.
94. C. Vallières, N. Fisher, T. Antoine, M. Al-Helal, P. Stocks, N. G. Berry, A. S. Lawrenson, S. A. Ward, P. M. O' Neill, G. A. Biagini and B. Meunier, *Antimicrob. Agents Chemother.*, 2012.

Chapter VI

Quantitative Structure Activity Relationships

6. Quantitative Structure Activity Relationships 302

6.1 Antimalarial 4-Aminoquinoline QSARs 303

6.1.1 Data Preparation 305

6.1.2 Molecular Descriptors 307

6.1.2.1 Objective Descriptor Selection Methods 308

6.1.2.1.1 Descriptor Correlation 308

6.1.2.1.2 CORCHOP 309

6.1.2.2 Subjective Descriptor Selection Methods 310

6.1.2.2.1 Forward Selection 310

6.1.2.2.2 Backward Elimination 311

6.1.2.2.3 Stepwise Regression 311

6.1.2.2.4 Genetic Algorithm 312

6.1.3 QSAR Development Excluding 3D Descriptors 313

6.1.3.1 Multiple Linear Regression 313

6.1.3.2 Internal Validation 314

6.1.3.2.1 Standard Error of Prediction 314

6.1.3.2.2 F-Statistic 315

6.1.3.2.3 t-Statistic 316

6.1.3.2.4 Cross-Validation 316

6.1.3.2.5 Bootstrapping 318

6.1.3.2.6 Y-Scrambling 318

6.1.3.2.7 Criteria for Internal Validation 319

6.1.3.3 External Validation 320

6.1.3.3.1 Sphere Exclusion Algorithm 321

6.1.3.3.2 Assessing Predictive Power 322

6.1.3.3.3 Criteria for External Validation 324

6.1.3.4 NF54 Dataset 324

6.1.3.5 K1 Dataset 328

6.1.4 QSAR Development Including 3D Descriptors 330

6.1.4.1 Conformation Search & Similarity Analysis 330

6.1.4.2 NF54 Dataset 333

6.1.4.3 K1 Dataset 338

6.1.5 Combinatorial MLR Calculations 343

6.1.6 Partial Least Squares 345

6.1.6.1 PLS Models 347

6.1.6.1.1 NF54 Dataset 348

6.1.6.1.2 K1 Dataset 349

6.1.7 Descriptor Frequencies 351

6.1.8 Descriptor Interpretations 351

6.1.8.1 Hy 352

6.1.8.2 JGI5 353

6.1.8.3 Mor31e 353

6.1.8.4 HARD 354

6.1.9 Descriptors and Chloroquine Resistance 354

6.1.10 *k*-Nearest Neighbour 355

6.1.10.1 *k*NN Models 357

6.1.11 Selective/Resistance Index 363

6.1.12 Summary of 4-Aminoquinoline QSAR Analysis 365

6.2 Hepatitis C Thiazolides QSARs 366

6.2.1 Hepatitis C Virus 366

6.2.2 Data Preparation 370

6.2.3 Model Development 371

6.2.4 Summary of Thiazolide QSAR Analysis 377

6.3 References 379

6. Quantitative Structure Activity Relationships

Quantitative structure activity relationships (QSAR) involve correlating the structural and/or physicochemical features of a molecule with a measured property, such as biological activity,[1] with most QSARs taking the general form of equation 6.1 (Chapter I).

$$\Delta Activity = f(\Delta Molecular\ descriptors)$$

Eq. 6.1 General QSAR equation.

The most popular and widely used machine learning technique for QSAR is that of MLR (eq. 6.2), which is an extension of simple linear regression.[1] The dependent variable (y) corresponds to the property of interest (i.e. biological activity), whilst the independent variables (x) are the molecular descriptors used. Statistical methods such as the squared correlation coefficient (r^2) can then be used to assess the performance of a model. However, there are many additional criteria which need to be satisfied in order to be confident of a models performance, as well as several other machine learning methods which can be considered to develop these QSARs.

$$y = m_1x_1 + m_2x_2 + \cdots m_nx_n + c$$

Eq. 6.2 Multiple linear regression equation.

In this thesis, QSAR methods have been successfully applied to two distinct datasets, the results of which will now be discussed. The first dataset was that of a number of 4-aminoquinoline analogues with reported antimalarial activities, whilst the other was a number of thiazolide compounds which had been tested against hepatitis C. QSAR methods and analysis procedures will be fully introduced as and when they are used.

6.1 Antimalarial 4-Aminoquinoline QSARs

4-aminoquinoline compounds (i.e. CQ) continue to attract interest because their mechanisms of action and resistance appear to be unrelated.[2] It is commonly accepted that CQ exerts its antimalarial activity by inhibiting haemozoin formation in the digestive vacuole of the parasite,[3-5] with the cause of resistance thought to be an increased efflux of the drug from the vacuole.[6, 7] However, several analogues and derivatives of CQ retain significant activity against CQR *P.falciparum* strains,[8-11] leading to suggestions that resistance could be compound specific, and not related to changes in the structure of the drug target.[2] Owing to this, there is still much potential within the 4-aminoquinoline class of compounds in attempting to combat or circumvent this resistance. QSAR is one computational approach which can be used to enhance our understanding of how certain chemical agents elicit their biological response. The results from QSAR can assist in optimising existing compounds in order to increase potency, with insight garnered later fed directly back into synthetic work.

The US patent *'Method of treating chloroquine-resistant malaria with aminoquinoline derivatives'*,[12] describes aminoquinoline derivatives of the general formula shown in figure 6.1. The R groups ranged from simple chlorine or hydrogen atoms, to carbon chains and trifluoromethyl groups (the appendices for this chapter are all on the disk in the stated folders; *Quantitative Structure Activity Relationships\Appendix 01*).

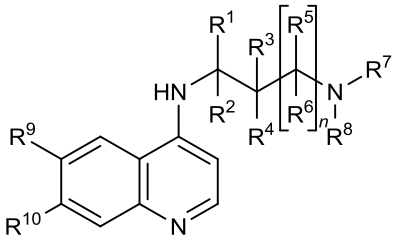


Fig. 6.1 Aminoquinoline template described in US patent 5596002. (W. Hofheinz, C. Jaquet and S. Jolidon, *United States Patent*, 1997, **5596002**.)

The compounds were all tested against a CQS (NF54) and CQR (K1) strain of the *P.falciparum* parasite, with values shown in table 6.1. The observed resistance between these two strains is most likely attributed to molecular differences in the parasites, as a result of point mutations. The activity of CQ was also recorded for reference.

Table. 6.1 IC_{50} values for the 4-aminoquinoline compounds from Patent 5596002 tested against CQS (NF54) and CQR (K1) strains of malaria.

ID	**CQS Strain** (NF54) IC_{50} (ng/ml)	**CQR Strain** (K1) IC_{50} (ng/ml)
1	4	9
2	7	14
3	7	12
4	7	15
5	4	7
6	5	9
7	4	8
8	5	11
9	11	32
10	7	17
11	5	18
11a	7	15
12	6	10
12a	7	9
13	30	47
14	9	21
15	2	6
16	3	9
17	3	15
18	4	14
19	7	9
19a	6	9

20	3	10
21	21	34
22	14	22
23	6	15
24	7	15
25	8	22
26	7	15
27	8	18
28	7	16
29	6	16
30	7	14
31	5	11
32	7	16
33	8	14
33a	7	8
33b	8	17
33c	8	24
33d	8	23
33e	12	53
33f	12	41
33g	5	17
33h	7	18
CQ	8	114

The dataset was a suitable candidate for QSAR analysis as models could be developed for the biological data of both strains, enabling a comparison between the models, and conclusions to be drawn with regard to potential resistance mechanisms. Individual QSAR models could also be interpreted.

QSARs were developed for both strains through a three stage process:[13]

i. Data preparation

ii. Data analysis

iii. Model validation

6.1.1 Data Preparation

The 45 compounds reported in the dataset were first constructed and energy minimised using the *'Energy Minimisation Protocol'* as described in the

Experimental Chapter. Next, the IC_{50} values for the NF54 and K1 strains were converted into activity values using negative log transformations to normalise the data,[14] and are reported in table 6.2 (*Quantitative Structure Activity Relationships\Appendix 02*). Two files were then created, one for each strain, with both containing the 45 structures together with their respective activity values (as per table 6.2). For both strains the activity values covered a sufficient range, with guidelines suggesting the spread of values across the dependent variable from the least active to the most active should be greater than two orders of magnitude for successful QSAR models.[15]

Table. 6.2 IC_{50} values and negative log transformations for the 45 compounds tested against CQS (NF54) and CQR (K1) strains of malaria.

ID	**CQS Strain** (NF54)		**CQR Strain** (K1)	
	IC_{50} (ng/ml)	Activity Value	IC_{50} (ng/ml)	Activity Value
1	4	7.82	9	7.47
2	7	7.62	14	7.32
3	7	7.62	12	7.38
4	7	7.64	15	7.31
5	4	7.80	7	7.55
6	5	7.74	9	7.49
7	4	7.84	8	7.54
8	5	7.76	11	7.42
9	11	7.48	32	7.02
10	7	7.67	17	7.29
11	5	7.80	18	7.25
11a	7	7.65	15	7.32
12	6	7.68	10	7.46
12a	7	7.60	9	7.49
13	30	7.01	47	6.81
14	9	7.55	21	7.18
15	2	8.12	6	7.64
16	3	7.99	9	7.51
17	3	7.98	15	7.29
18	4	7.88	14	7.34
19	7	7.70	9	7.59
19a	6	7.66	9	7.49
20	3	8.03	10	7.50
21	21	7.17	34	6.96

22	14	7.31	22	7.11
23	6	7.70	15	7.31
24	7	7.64	15	7.31
25	8	7.60	22	7.16
26	7	7.64	15	7.31
27	8	7.60	18	7.25
28	7	7.66	16	7.30
29	6	7.72	16	7.30
30	7	7.64	14	7.34
31	5	7.82	11	7.48
32	7	7.67	16	7.31
33	8	7.54	14	7.29
33a	7	7.57	8	7.51
33b	8	7.54	17	7.21
33c	8	7.52	24	7.04
33d	8	7.52	23	7.06
33e	12	7.39	53	6.74
33f	12	7.39	41	6.85
33g	5	7.72	17	7.19
33h	7	7.58	18	7.17
CQ	8	7.60	114	6.45

6.1.2 Molecular Descriptors

Molecular descriptors (Chapter I) are calculated based on a molecular structure, and encode chemical and/or physicochemical information in numerical form.[16] They are generated using specific programs which perform various feature recognition and transformation steps, to generate a series of values for each compound.[17, 18] One such program is DRAGON 3.0[19] which encodes almost 1500 descriptors (1497), encompassing 0, 1, 2 and 3D molecular properties. Another program is ADMEWORKS Modelbuilder,[18] which can describe just over 500 descriptors (523).

Through combining 0, 1, 2 and 3D molecular descriptors, it is hoped that the bulk of a molecules properties will be fairly well represented. However, the generation of 3D descriptors may be problematic, requiring the prediction of a compounds active conformation, which can be difficult for highly flexible molecules.[20] With this in

mind, QSAR analysis can still prove useful using just 0, 1 and 2D descriptors,[21] providing reliable and interpretable results that are easily automated and adapted to the task of database searching, or virtual screening.

Given the large number of molecular descriptors available, selection procedures can be readily used to identify those descriptors which best represent the data. It is only natural that given the large amount of information there is likely to be some redundancy, therefore descriptor selection is an integral part of QSAR development, helping to reduce the often vast number of descriptors, to increase the chance of finding a significant QSAR model.[22]

6.1.2.1 Objective Descriptor Selection Methods

Descriptor selection methods are divided into two categories; objective or subjective. Objective selection methods involve removing descriptors based on their relationship to the other descriptors, as it is ultimately important to consider only those descriptors which describe unique information about the chemical structures, removing those which are highly correlated to others.

6.1.2.1.1 Descriptor Correlation

Descriptor correlations should be checked as a matter of routine to avoid over representing a dataset. Many correlations can be identified from simple scatterplots between pairs of descriptors, with the ideal situation being no discernible relationship between the two. However, when many descriptors need to be considered it is much more convenient to compute a pair wise correlation matrix, one that quantifies the degree of correlation between all pairs of descriptors. Each entry (i,j) in the correlation matrix represents the correlation coefficient between the

descriptors x_i versus x_j. The correlation coefficient r is then found using equation 6.3, where N is the number of molecules, k is set to 1, and $\langle x \rangle$ is the mean of the independent variables.[1]

$$r = \frac{\sum_{k=1}^{N}[(x_{i,k} - \langle x_i \rangle)(x_{j,k} - \langle x_j \rangle)]}{\sqrt{\sum_{k=1}^{N}(x_{i,k} - \langle x_i \rangle)^2 \sum_{k=1}^{N}(x_{j,k} - \langle x_j \rangle)^2}}$$

Eq. 6.3 Correlation coefficient equation.

The value of the correlation coefficient ranges from -1.0 to +1.0. A value of +1.0 indicates a perfect positive correlation, with a plot of x_i versus x_j giving a straight line with a positive gradient. Conversely, a value of -1.0 indicates a perfect negative relationship. A gradient of zero would suggest that no relationship exists between the variables, and it is these descriptors which best produce the more meaningful QSAR models.

When a pair of descriptors are found to be highly correlated, a decision needs to be made as to which to keep and which to discard. One approach is to remove the descriptor which has the most correlations to other descriptors, yet one may also choose to keep this descriptor, in order to reduce the size of the dataset.

6.1.2.1.2 CORCHOP

CORCHOP provides a means for making such decisions. With the growing number of descriptors used to characterise molecules, the amount of information available can be overwhelming, so it is important to extract the information which best represents the data, and avoid "chance" correlations. Although techniques to analyse correlations have been well implemented, there are issues when deciding which highly correlated parameters to discard. The computer routine CORCHOP was

devised to aid in the systematic reduction of large amounts of data, by reducing the number of descriptors whilst retaining the intrinsic information they describe.[23] When two descriptors each correlate with the same descriptors, it is essential that one be removed. CORCHOP acts by removing those descriptors whose distribution deviates most from normal.

6.1.2.2 Subjective Descriptor Selection Methods

With highly correlated descriptors removed, subjective selection methods can be used to select the most appropriate descriptors, based on their relationships to the dependent variable. When the number of descriptors in a dataset is very large compared to the number of compounds it contains, a learning algorithm is faced with the problem of selecting a relevant subset of these descriptors, from which to develop the QSAR models.[24] A general rule of thumb is that there should be at least five compounds per descriptor in an MLR regression.[1] Several subjective selection methods will now be introduced.

6.1.2.2.1 Forward Selection

Forward selection is a searching method that begins with an empty set of features,[24] and generates an equation containing just one descriptor from a dataset, typically chosen to be the variable which would contribute most to a model, as assessed by the t-statistic (see later).[1] N models are learnt containing just one descriptor each, with the best model chosen based on its statistics, effectively selecting the best descriptor to model the given property. Subsequent descriptors are then added using the same criteria. $N-1$ feature subsets are made by pairing the chosen descriptor with all the remaining $N-1$ descriptors, one by one. $N-1$ models are learnt and their

statistics compared to select the best performing model, in effect, selecting the next best descriptor to explain the dependent variable. The process is repeated and terminates when either a predetermined number of variables have been reached, or when the last variable has an insignificant contribution to the regression equation.[25]

6.1.2.2.2 Backward Elimination

Backward elimination refers to a search involving the full set of descriptors,[26] and starts with an equation using all N descriptors, each of which is then periodically removed in turn, usually starting with the descriptor that has the smallest contribution to the reduction of predictive sum of squares (see later),[25] or which has the smallest t-statistic.[1] Each descriptor is then dropped one by one, and N models are learnt containing $N-1$ descriptors. Based on the statistics of these N models, the best model is selected to identify the best feature set, effectively eliminating the worst descriptor for modelling the given dependant variable. $N-1$ feature subsets containing $N-2$ descriptors are then developed by dropping all of the descriptors one by one. $N-1$ models are learnt and their statistics compared to select the best performing model. The best is then selected, eliminating the next least important descriptor. These iterations are repeated until a predefined target size is reached, or until all variables are considered significant. A comparison between forwards and backwards selection found that forward selection generally performs better.[24]

6.1.2.2.3 Stepwise Regression

In a stepwise regression a variable that enters a model may in the earlier stages of selection be deleted.[25] It is essentially a forward selection routine, but at each stage the possibility of deleting a variable (as with backward elimination) is considered.

The number of variables retained in a model is based on the levels of significance assumed for inclusion and exclusion of variables from the model. The model building terminates when no variable in the new model can be removed, and all the next best candidates cannot be retained.

6.1.2.2.4 Genetic Algorithm

Genetic algorithms (GAs) are based on various computational models of Darwinian evolution,[27] and have found use in data analysis to reduce the number of features in a regression model. They can be used to generate a population of linear regression QSAR equations, each with different combinations of descriptors.[28] Over a number of generations, these equations undergo mutation, with some being removed (death rate), some new ones generated (birth rate), and the best ones being carried on from one generation to another (elitism). The output from this is a family of models from which one can select the highest scoring model, or analyse the data further to determine the most commonly occurring descriptors.[1, 22]

GAs do not search for just one solution, but instead a population, attempting to increase the fitness of the population at each generation.[22] Each solution is evaluated based on some domain-specific function, then kept or discarded based on that evaluation. If discarded, that member of the population is replaced by a new solution, which is created by combining parts of good solutions. This process is repeated over and over, combining different aspects of good solutions, searching for an optimal combination of features.

6.1.3 QSAR Development Excluding 3D Descriptors

An array of descriptor selection methods have now been discussed, but in order for them to be applied, descriptor subsets had to be calculated for the compounds contained within the aminoquinoline datasets. To begin with, QSARs were developed using only the 0, 1 and 2D descriptors. This was done to remove the initial need for conformational analysis and alignment of the compounds to a common pharmacophore, required for the calculation of 3D descriptors. For the 45 structures a total of 957 descriptors were calculated using both ADMEWORKS Modelbuilder[18] and DRAGON 3.0.[19] QSARs were built using the program PHAKISO[29] (Pharmacokinetics *In Silico*) which contains a host of functions and machine learning methods highly amenable to model development. These algorithms allow computers to learn and recognise complex patterns, making intelligent decisions based on inputted data. [30, 31]

6.1.3.1 Multiple Linear Regression

MLR has proven to be a very useful tool for the generation of QSAR models.[1] However, before relationships such as those in equation 6.2 could be developed, it was crucial that the descriptors be normalised. Molecular descriptors were therefore autoscaled (Chapter I), meaning that the variables each had a mean of zero and a standard deviation of one.[1] MLR attempts to find the smallest possible sum of squared differences between the actual dependent observations, and those predicted from the regression equation. The most common method of assessing a regression equation is the squared correlation coefficient (r^2) (Chapter I). It is calculated using equation 6.4, and allows comment to be drawn as to the estimated proportion of the variation in the dependent variable that is explained by the regression equation.[30]

When $r^2 = 0$ then there is no linear relationship observed, whilst values greater than zero suggest a stronger relationship between the two, up to a maximum value of one, indicating a perfect fit.

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS}$$

Eq. 6.4 Calculation of the r^2 relationship.

The literature suggests that a good model has an r^2 value greater than 0.7, but less than one, as a value of one is usually indicative of over training.[31, 32] However, whilst r^2 is useful in highlighting promising models, several other statistical parameters should also be considered in order to be certain of its validity, and thus predictive potential.

6.1.3.2 Internal Validation

r^2 is part of a family of statistical methods that fall under the umbrella of internal validation. Internal validation techniques are used to assess the performance of a model to see how well it predicts the dependant variable for the set of molecules which were used to develop the equation.[33, 34] These techniques can be used to overcome some of the limitations of the r^2 statistic alone, mainly in cases where the data has been over fitted (i.e. $r^2 = 1$). Generally however, the r^2 value for a model can be increased with the addition of more independent variables, even if the added variable does not reduce the unexplained variance of the dependent variable.

6.1.3.2.1 Standard Error of Prediction

The standard error of prediction (s) indicates how well the regression function predicts the observed data, and is calculated by equation 6.5.[1] RSS (eq. 1.12)

represents the residual sum of squares, n the number of data points, and p the number of independent variables in the equation.

$$s = \sqrt{\frac{RSS}{n - p - 1}}$$

Eq. 6.5 Calculation of standard error of prediction.

6.1.3.2.2 F-Statistic

The Fisher statistic (F), or the variance ratio, is another commonly used measure to assess a linear regression, calculated by dividing ESS (eq. 1.11) by RSS (eq. 1.12), as shown in equation 6.6. [1, 22] The form of this equation reflects the number of degrees of freedom associated with each parameter.

$$F = \frac{ESS}{p} \frac{N - p - 1}{RSS}$$

Eq. 6.6 Calculation of the Fisher statistic.

The F-statistic is used to test the statistical significance of the observed differences amongst the means of two or more random samples. The experimental value of F can be compared with values from statistical tables, or those calculated manually at different confidence levels. If the experimental value is greater than the tabulated value, then the equation is said to be significant at that particular confidence level. Higher values of F therefore suggest a higher level of significance, increasing the confidence in the models ability. At a given confidence level the threshold value of F falls as the number of independent variables decreases, and/or the number of data points increases, consistent with the desire to describe as large a number of data points, with as few independent variables as possible.

6.1.3.2.3 t-Statistic

The t-statistic comments on the significance of each independent variable in the linear regression equation.[1] If k_i represents the coefficient in the regression equation associated with a particular variable x_i, then the t-statistic is given by equation 6.7, where $s(k_i)$ is the standard error of the coefficient.

$$t = \left|\frac{k_i}{s(k_i)}\right|; a(k_i) = \sqrt{\frac{RSS}{N-p-1}.\frac{1}{\sum_{j=1}^{N}\left(x_{i,j} - \langle x_i \rangle\right)^2}}$$

Eq. 6.7 Calculation of the t-statistic.

Similar to the F-statistic, values of t can be compared to those listed in statistical tables at various confidence levels. If the calculated value is greater than that in the table, then the coefficient is considered significant at that confidence level.

6.1.3.2.4 Cross-Validation

Cross-validation is one of the most common validation techniques. It involves the removal of one or more compounds form a dataset, in such a way that each object is removed only once. A QSAR model is then derived given the remaining data, which is then used to predict the dependent variable of the molecule/s removed. The squared differences between the true response (y_i) and the predicted response ($y_{pred,i}$) for each compound left out are added to give the predictive residual sum of squares (PRESS), as calculated by equation 6.8. PRESS is analogous to RSS, but rather than using the calculated values ($y_{calc,i}$) from the model, PRESS uses the predicted values ($y_{pred,i}$) for data not used to derive the model.

$$PRESS = \sum_{i=1}^{N} \left(y_i - y_{pred,i}\right)^2$$

Eq. 6.8 Calculation of the predictive residual sum of squares.

The simplest cross-validation procedure is the leave-one-out (LOO) approach, where each compound is removed one at a time. The process is repeated for every entry, leading to the cross-validated r^2 statistic (q^2), calculated by equation 6.9. q^2 is always lower than r^2, and whilst r^2 measures the goodness-of-fit, q^2 attempts to measure the goodness-of-prediction.

$$q^2 = 1 - \frac{PRESS}{TSS}$$

Eq. 6.9 Calculation of q^2.

The leave-many-out (LMO) cross-validation approach can also be used, and may provide a more realistic estimate as to the predictive ability of a model, particularly for larger datasets. The dataset can be divided into several groups, each of which is left out in turn to generate the q^2 statistic. By repeating this procedure a large number of times, selecting different groupings at random each time, a mean q^2 can be derived. For a well-behaved system there should not be much variation in the equations obtained during such a procedure, nor should the r^2 value calculated using the entire dataset be significantly larger than the mean q^2 value. If there is a large discrepancy, then it is likely that the data has been over-fitted, and the model may not be robust.

Often a high value of q^2 (greater than 0.5) is considered proof of a models predictive ability, but this assumption can be misleading.[35] A high q^2 does not automatically mean that the model has a high predictive ability, as the only way to estimate its true predictive power is to externally validate it on a test set. A test set represents a

subset of compounds that were not used during model development, and to which a model can be applied. A high q^2 value is therefore a necessity for model validation, but is not sufficient in itself.[35]

6.1.3.2.5 Bootstrapping

Bootstrapping is a widely popular technique that can provide answers when other validation procedures fail.[22] It is based on building a distribution sample for a statistic by re-sampling the initial data.[36] The idea is to simulate what would happen if the samples were randomly selected from a dataset which was representative of the entire population.

Typically, K groups of size n are generated by a repeated random selection of n objects from the original dataset.[37] Some of these objects can be included in the same random sample several times, whilst others will never be selected. The model obtained on the n randomly selected objects is used to predict the target properties for the excluded sample. The difference between the parameters calculated from the original dataset and the average of the parameters calculated from the bootstrap samplings is a measure of the bias of the original calculation. The calculated variance of the parameter estimates the accuracy with which any of the parameters can be estimated from the input data, and as with the cross-validation technique, a high average q^2 in bootstrap validation is indicative of a robust model.[37]

6.1.3.2.6 Y-Scrambling

Y-scrambling can also prove to be another useful internal validation method.[38] It is widely used to validate the robustness of a QSAR model, identifying models based on chance correlation.[22] When selecting descriptors that are of relevance to a

particular property, it is possible to find some descriptors purely by chance. Using correlation-based selection methods it is possible to select descriptors that are of no real significance, but seem to fit the property of interest well. Y-scrambling attempts to identify these chance correlations.

Y-scrambling is performed by calculating the quality of models, via r^2 and q^2, when the dependent variable has been randomly modified. It is modified by randomly assigning each compound a different dependent variable from the true set of responses.[39] If the original model is genuine, then there is a significant difference in the quality of the original model and that found using Y-scrambling. This procedure is repeated several hundred times, and it is expected that the resulting QSAR models should generally have low r^2 and q^2 values.[37]

6.1.3.2.7 Criteria for Internal Validation

Table 6.3 illustrates a list of internal validation parameters which were compiled to determine the success of a QSAR model. These guidelines were based on a consensus across the literature, and were used to assess the developed models.[1, 13, 22, 35, 40, 41]

Table. 6.3 Internal validation criteria.

Parameter	Threshold Value
General Parameters	
Number of data points	≥ 10
Dependent variable data range	≥ Two orders of magnitude
Molecules/Descriptor Ratio	≥ 5
Internal Validation Statistics	
r^2	> 0.7
q^2	> 0.5
F-statistic	Higher than table value
Bootstrapping	No more than 0.3 lower than r^2
Dependent Variable Statistics	
t-statistic	≥ 2

Though internal validation is essential, it has been recommended that the only true test of a model lays in its predictive ability,[13] which can be assessed using external validation. For this there exists an additional set of criteria.

6.1.3.3 External Validation

External validation greatly influences the success of a model, as it is highly desirable to be able to accurately predict the activity of compounds which were not used during model development. Though internal validation allows for comment to be drawn as to the robustness of a particular model, it gives no rational measure as to its predictive ability. Therefore, external validation is the only true predictive test for a model, and provides supplementary information in support of internal validation.[37]

It has been demonstrated that the widely accepted q^2 statistic is an inadequate characteristic with which to assess the predictive capacity of a model.[35] There is also evidence to support that there exists no correlation between the values of q^2 for the training set, and the accuracy of prediction (r^2) for the test set.[13] The only way to estimate the true predictive power of a model is to test it on a sufficiently large collection of compounds from an external test set. There is typically insufficient data available to test newly synthesised compounds, so the best approach is to use a rationally selected training and test set, chosen from the original data.[13] It has been argued that training and test sets must satisfy the following criteria:[42, 43]

i. Representative points of the test set must be close to those of the training set
ii. Representative points of the training set must be close to representative points of the test set
iii. Training set must be diverse

It has been demonstrated that models found using a rational division of an experimental dataset, generally lead to better predictive models.[42] There are several approaches available to split datasets.

6.1.3.3.1 Sphere Exclusion Algorithm

The sphere-exclusion algorithm is one such approach, and is widely used for the comparison of chemical databases or libraries.[44-46] It attempts to specify which compounds most effectively cover the available property space, with the basic sphere-exclusion algorithm taking the following form:[1]

1. Define a threshold dissimilarity, t
2. Select a compound and place it in the subset
3. Remove all molecules from the dataset that have a dissimilarity to the selected molecule of less than t
4. Return to step 2 if there are molecules remaining in the dataset

The first compound is selected for inclusion in the subset, after which all compounds in the dataset that are within the similarity threshold to the selected compound are removed from consideration. This is analogous to enclosing the compound within a hypersphere and removing all compounds from the dataset that fall within the sphere. This sphere can be constructed using equation 6.10, where V is the total volume occupied by the representative points in the normalised descriptor space, K the number of descriptors, N the number of molecules, and c the dissimilarity level, which can be varied to construct different training and test sets.[45, 46] Several variants of the algorithm exist, differing in the way in which the first compound is selected, the threshold value used, and the way that the next compound is selected at each iteration.

$$R = c(V/N)^{1/K}$$

Eq. 6.10 Sphere exclusion algorithm.

6.1.3.3.2 Assessing Predictive Power

Once a suitable training and test set split is identified models can be developed and validated. It is the ability of the QSAR model to accurately predict the dependent variable of the test set which really assesses its predictive capabilities, and is quantitatively estimated by the external q^2 statistic, calculated using equation 6.11. [37] y_i represents the observed response for the i^{th} object, $\hat{y}_i$ the predicted response for the i^{th} object, and $\bar{y}_{tr}$ the average response value for the entire training set.

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{tr})^2}$$

Eq. 6.11 Calculation of external q^2.

Further statistical parameters referred to as the Tropsha parameters, act to assess the predictive capabilities of a model, and are calculated based on graphical plots of the predictions.[35, 38] For an ideal model, the slope of the gradient when the observed and predicted dependent variables are plotted against one another should be equal to one, with an intercept of zero, and a correlation coefficient (R) of one. R can be calculated using equation 6.12, where $\bar{y}$ and $\hat{\bar{y}}$ are the average values of the observed and predicted activities respectively.

$$R = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})}{\sqrt{\sum(y_i - \bar{y})^2(\hat{y}_i - \hat{\bar{y}})^2}}$$

Eq. 6.12 Calculation of correlation coefficient.

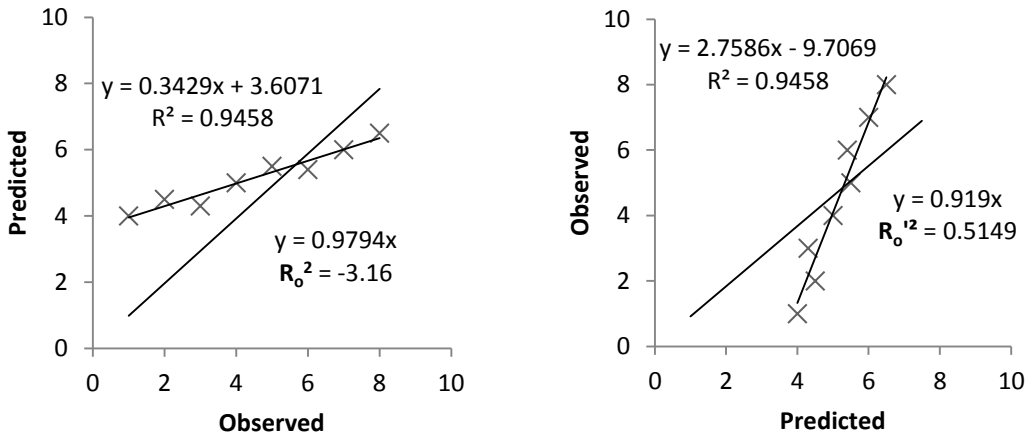
A real QSAR model could be considered to have a high predictive ability if it is close to the ideal situation. This implies that the correlation coefficient between the actual (y) and predicted ($\hat{y}$) activities must be close to one. Regressions of y against

$\hat{y}$ or $\hat{y}$ against y through the origin, i.e. $y^{r_0} = k\hat{y}$ and $\hat{y}^{r_0} = k'y$, respectively, should therefore be characterised by at least either k or k' being close to one, with the slopes k and k' calculated using equation 6.13.

$$k = \frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i^2} \qquad k' = \frac{\sum y_i \hat{y}_i}{\sum y_i^2}$$

Eq. 6.13 Calculation of the slope gradients for regression equations passing through the origin.

Plotting both y against $\hat{y}$ and $\hat{y}$ against y may appear redundant, however, these two plots can be distinguished by different statistics.[47] Figure 6.2 illustrates a case where the correlation between the actual activities and those predicted by a QSAR model for an external test set, gave an r^2 of 0.95, and k and k' values both close to 1, but yet despite this the predictions were found to be inaccurate. This was shown via regression lines through the origin, as defined by $y^{r_0} = k\hat{y}$ and $\hat{y}^{r_0} = k'y$ (intercepts set to zero).



(a) (b)

Fig. 6.2 Examples of regression between observed vs. predicted (a) and predicted vs. observed (b) activities for compounds from an external test set. Despite high r^2 and k and k' close to 1, the model is not highly predictive due to the regression lines through the origin not being close to the optimal regressions, and r_0^2 and $r_0'^2$ being considerably different to one another.

Both correlation coefficients for these lines (r_0^2 and $r_0'^2$), as calculated by equations 6.14, had very different values, and were also significantly different from that of r^2.

$$r_0^2 = 1 - \frac{\sum(\hat{y}_i - y_i^{r_0})^2}{\sum(\hat{y}_i - \hat{\bar{y}})^2} \qquad r_0^2 = 1 - \frac{\sum(\hat{y}_i - y_i^{r_0})^2}{\sum(\hat{y}_i - \hat{\bar{y}})^2}$$

Eq. 6.14 Calculation of r_0^2 and $r_0'^2$.

This example demonstrates the need to impose additional conditions to validate the predictive ability of a QSAR model. Another consideration is that the r^2 value should be sufficiently similar to that of r_0^2 and $r_0'^2$, with the difference between the two not being too high.

6.1.3.3.3 Criteria for External Validation

The external validation criteria have been compiled in table 6.4, based on observations across the literature.[35, 37, 41]

Table. 6.4 External validation criteria.

Parameter	Threshold Value
$\boldsymbol{q^2}$	> 0.5
$\boldsymbol{r^2}$	> 0.6
$\frac{(\boldsymbol{r^2 - r_o^2})}{\boldsymbol{r^2}}$	< 0.1
Slope of line through origin ($\boldsymbol{k}$)	$0.85 \leq k \leq 1.15$
$\lvert \boldsymbol{r_o^2 - r_0'^2} \rvert$	< 0.3

If a QSAR equation is found to satisfy both the internal and external validation criteria, then it is likely that a significant and valid model has been developed.

6.1.3.4 NF54 Dataset

QSAR models were first built for the NF54 dataset of compounds. *'QSAR Protocol 1'* as described in the Experimental Chapter, was employed to develop a number of models (*Quantitative Structure Activity Relationships\Appendix 03*) using all 45 compounds and excluding the 3D descriptors (957 descriptors used in total). Following objective descriptor selection, subjective methods were used to find the best subset of these descriptors in order to model the dependent variable. Backward

elimination and stepwise regression failed to produce any descriptor subsets that gave valid models, with forward selection also performing poorly, with the best model having an internal r^2 value of only 0.50. GA produced the best model for the data, which had an r^2 of 0.80, as shown in table 6.5.

Table. 6.5 Statistics for model 1.

Model 1

Subjective selection method	Genetic algorithm
Training set size	45
Number of descriptors	9
Machine learning method	MLR
Molecule/Descriptor Ratio	5

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.80	Yes
q^2	0.66	Yes
F-statistic	15.07	Yes (Table value 2.16)
Bootstrapping	0.42	Almost (Ideally > 0.50)

Despite a slightly low bootstrapping value the model suggested that the data could be modelled successfully, giving good internal statistics. However, the true test for the model would be its predictive ability.[13] The dataset was split using the sphere exclusion algorithm, with 37 molecules in the training set, and 8 in the test set. A general rule of thumb is that there should be between 15 to 40% of the total number of molecules in the test set.[48] *'QSAR Protocol 2'* as described in the Experimental Chapter was employed to build a model using GA subjective descriptor selection (*Quantitative Structure Activity Relationships\Appendix 03*), the statistics for which are illustrated in table 6.6.

Table. 6.6 Statistics for model 2.

Model 2

Subjective selection method	Genetic algorithm
Training set size	37
Test set size	8
Number of descriptors	7
Machine learning method	MLR
Molecule/Descriptor Ratio	5.286

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.80	Yes
q^2	0.69	Yes
F-statistic	16.04	Yes (Table value 2.35)
Bootstrapping	0.61	Yes (Ideally > 0.50)
External Validation		
Parameter	**Value**	**Valid?**
q^2	-0.51	No
r^2	0.17	No
$(r^2 - r_o^2)/r^2$	0.34	No
k	0.99	Yes
$\left\|r_o^2 - r_o'^2\right\|$	1.21	No

The model was internally significant ($r^2 = 0.8$) but failed externally ($r^2 = 0.17$). Figure 6.3 illustrates the predicted versus actual activity values for the training and test sets, with a clear linear relationship observed for the training set, but not the test set, as an outlier appeared to skew the data. So whilst internally the model appeared to be robust and have good predictive capabilities (i.e. $q^2 = 0.69$), this was clearly not the case, emphasising the importance of external validation.

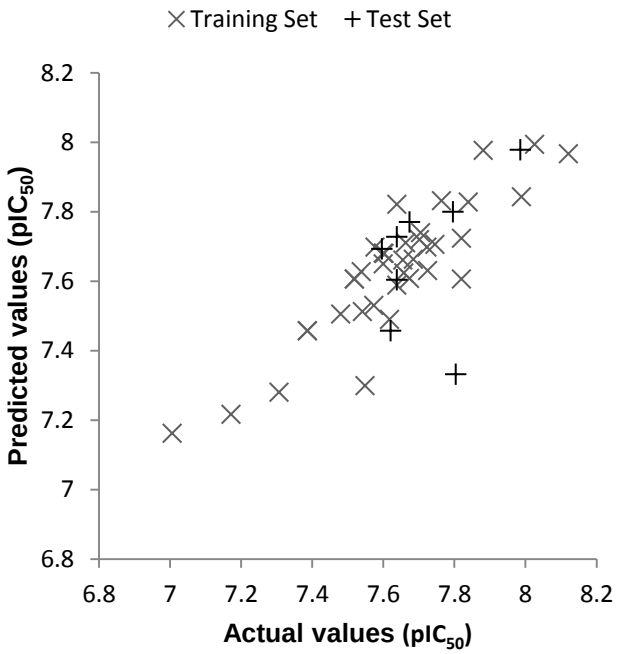
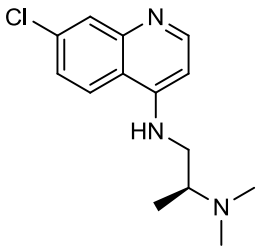


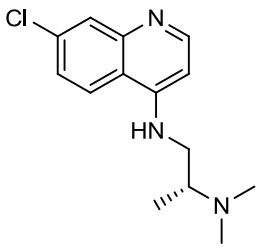
Fig. 6.3 Graph representing the predicted vs. actual activity values for model 2.

The reason for the poor predictive capabilities of the model were thought to lie with the descriptors used to describe the spread of molecular properties. There were several pairs of enantiomers in the dataset, with both compounds in each pair reporting different activities to one another. Therefore, the 0, 1 and 2D molecular descriptors were not sufficient to make this distinction, as they did not encode for stereocentres. Take for example molecules 33g and 33h (fig. 6.4). Both had the same structure but were enantiomers of each other, with 33g (*S* isomer) being more potent than 33h (*R* isomer) by a few ng/ml depending on the strain. The stereochemistry of these molecules plays a critical role in their activity, and represents their only distinguishing feature. 3D molecular descriptors are therefore essential to make this distinction.



33g

NF54 IC_{50} = 5 ng/ml (pIC_{50})
K1 IC_{50} = 17 ng/ml (pIC_{50})



33h

NF54 IC_{50} = 7 ng/ml (pIC_{50})
K1 IC_{50} = 18 ng/ml (pIC_{50})

Fig. 6.4 Molecules 33g and 33h.

6.1.3.5 K1 Dataset

Prior to the calculation of 3D descriptors, QSAR models were first developed for the K1 dataset using the same 0, 1 and 2D descriptors (957 descriptors in total), mainly to support the hypothesis that 3D descriptors were essential to accurately model activity. *'QSAR Protocol 1'* was used to develop regression models (*Quantitative Structure Activity Relationships\Appendix 04)*, with GA subjective descriptor selection once again providing the best model ($r^2 = 0.73$), as shown by table 6.7.

Table. 6.7 Statistics for model 3.

Model 3

Subjective selection method	Genetic algorithm
Training set size	45
Number of descriptors	9
Machine learning method	MLR
Molecule/Descriptor Ratio	5

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.73	Yes
q^2	0.43	No
F-statistic	10.39	Yes (Table value 2.16)
Bootstrapping	-9.57	No (Ideally > 0.43)

Using only 0, 1 and 2D descriptors no models were found that passed all of the internal validation criteria, with model 3 having a q^2 value of only 0.43. Again, this is most likely due to insufficient molecular descriptors to model activity and

distinguish between closely related compounds. For completeness, models were developed and tested externally using the K1 data via *'QSAR Protocol 2'* (*Quantitative Structure Activity Relationships\Appendix 04*), with the statistics for the best model shown in table 6.8.

Table. 6.8 Statistics for model 4.

Model 4

Subjective selection method	Genetic algorithm
Training set size	37
Test set size	8
Number of descriptors	7
Machine learning method	MLR
Molecule/Descriptor Ratio	5.286

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.80	Yes
q^2	0.70	Yes
F-statistic	16.95	Yes (Table value 2.35)
Bootstrapping	-3.05	No (Ideally > 0.50)
External Validation		
Parameter	**Value**	**Valid?**
q^2	-0.2.44	No
r^2	0.246	No
$(r^2 - r_o^2)/r^2$	0.09	Yes
k	0.99	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	5.65	No

Though internally this model saw an improvement over model 3, externally it was poorly predictive. The improvement internally may have been due to the smaller training set, or perhaps the sphere exclusion algorithm simply split the data in a manner better suited for regression modelling,[13] removing the most dissimilar compounds. Regardless, the model still failed externally. Figure 6.5 illustrates more clearly this failure, with the data of the test set at odds with the strong linear relationship observed by the training set.

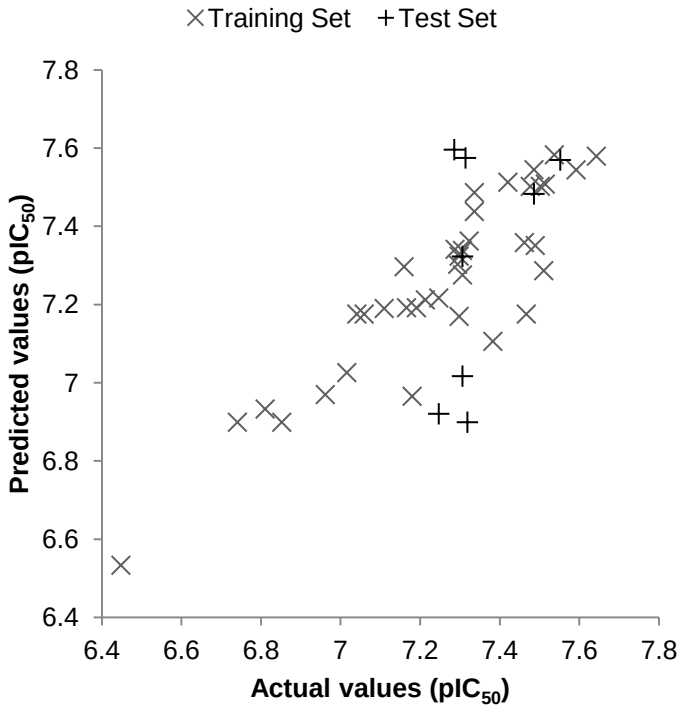


Fig. 6.5 Graph representing the predicted vs. actual activity values for model 4.

Though the 0, 1 and 2D molecular descriptors could produce internally significant models for both datasets, externally they showed little potential as predictive entities. 3D molecular descriptors were therefore investigated in attempts to overcome the limitations of these models.

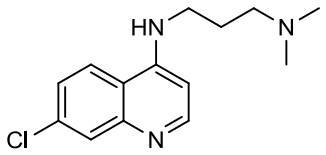
6.1.4 QSAR Development Including 3D Descriptors

To calculate the 3D descriptors it was first necessary to map each of the compounds to a known pharmacophore or bioactive conformation from across the series. This was done using conformation search and similarity analysis methods.

6.1.4.1 Conformation Search & Similarity Analysis

The pharmacophore/bioactive conformation of the 45 compounds was taken to be the lowest energy conformation of the most active compound in the series. Usually the lowest energy confirmation fits the pharmacophore,[49] due to low energy

conformers being more highly popular than others according to Boltzmann distribution.[50] Compound **15** (fig. 6.6) was found to have the most potent activity against both parasite strains, therefore a conformer distribution calculation of its structure was performed using the *'Conformer Distribution Protocol'* detailed in the Experimental Chapter. 100 conformations were calculated, with the energy (kJ/mol) and Boltzmann distribution values for several of the most promising conformations reported in table 6.9. The bioactive conformation was identified as that which had the lowest energy/highest Boltzmann distribution values (conformer 001).



15

Fig. 6.6 Molecule **15**.

Table. 6.9 Energy and Boltzmann distribution values for several conformers of molecule **15**.

Conformer	Boltzmann Distribution	Energy (kJ/mol)
001	0.774661687	139.319048
013	0.067193963	145.38013
024	0.062738299	145.550227
035	0.010870281	149.89599
046	0.009892085	150.129766
057	0.007285027	150.888166
068	0.007183015	150.923127
079	0.005686303	151.502397

The chemical function descriptors (CFDs) available within Spartan '08[51] were used to define the pharmacophore, and are described in table 6.10. Figure 6.7 illustrates the lowest energy conformation of compound **15** (hydrogen atoms not shown), together with labels indicating the corresponding CFDs. It is important to note that Spartan '08[51] does not consider the directionality of the pharmacophore, such as the direction of H-bond donors or acceptors, simply the presence of absence of particular features.

Table. 6.10 Chemical functional descriptors.

Label	CFD	Chemical Meaning
1	Hydrophobe	Sterically-crowded region
2	Aromatic	Aromatic π system
3	Hydrogen-bond donor	Acidic hydrogen
4	Hydrogen-bond acceptor	Lone pair
5	Positive ionisable site	Basic site
6	Negative ionisable site	Acidic site

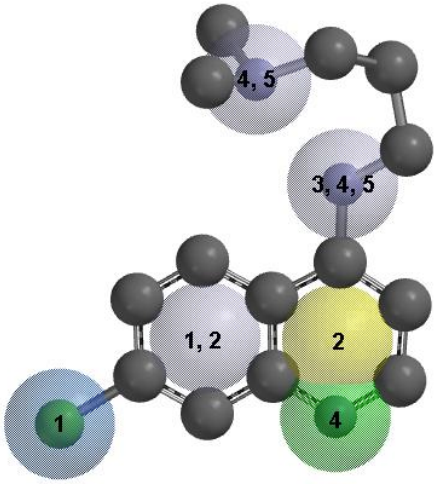


Fig. 6.7 CFDs of the lowest energy conformation of compound **15** (pharmacophore).

A conformer library was generated for the other 44 molecules in the dataset, using the *'Conformer Library Protocol'* as described in the Experimental Chapter. The conformer of each compound that most closely matched the pharmacophore was assumed to be that compounds bioactive conformation. The CFD pharmacophore of compound **15** was first defined and then each conformer in the library for all molecules compared to it. This was done using the *'Similarity Analysis Protocol'* as described in the Experimental Chapter. Each conformer was scored between zero and one, with a score of one indicating perfect similarity between the conformer and the pharmacophore (according to CFDs), and a score of zero indicating no similarity.[52] The highest scoring conformer for each molecule (all of which had a similarity score above 0.6) was taken to be the bioactive conformer for that compound.

673 3D descriptors were calculated with DRAGON 3.0[19] using the bioactive conformation of the 45 compounds, making for a total of 1630 descriptors, describing the 0, 1, 2 and 3D chemical properties of the molecules. QSAR models could then be developed to see how the inclusion of 3D descriptors affected the results.

6.1.4.2 NF54 Dataset

With this increased number of descriptors, a different approach was taken to reduce these to only those which best described the activity. To begin, models were developed for the NF54 dataset using all descriptors via *'QSAR Protocol 2'* (*Quantitative Structure Activity Relationships\Appendix 05*), with table 6.11 describing the most successful model.

Table. 6.11 Statistics for model 5.

Model 5

Subjective selection method	Genetic algorithm
Training set size	40
Test set size	5
Number of descriptors	8
Machine learning method	MLR
Molecule/Descriptor Ratio	5

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.84	Yes
q^2	0.75	Yes
F-statistic	20.47	Yes (Table value 2.26)
Bootstrapping	0.66	Yes (Ideally > 0.54)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.56	Yes
r^2	0.72	Yes
$(r^2 - r_o^2)/r^2$	0.002	Yes
k	1.004	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	0.14	Yes

This model (model 5) had excellent internal and external statistics, but the sphere exclusion algorithm only allowed a split with a maximum of five compounds in the test set, despite alterations to the dissimilarity parameters. This is most likely due to the large number of descriptors and difficulties in dividing the data. To combat this it was necessary to reduce the number of descriptors prior to splitting. The 1630 descriptors were calculated using ADMEWORKS Modelbuilder[18] (523 descriptors) and DRAGON 3.0[19] (1107 descriptors), so models were built for each of these descriptor sets independently, using *'QSAR Protocol 1'* (*Quantitative Structure Activity Relationships\Appendix 06*). The descriptors from each of the valid models were inspected, and those with a t-statistic greater than 2 were considered to be the best in describing the activity of the NF54 strain. In total, 38 descriptors across the two descriptor sets obeyed this criterion. This subset of descriptors were used to develop further models according to *'QSAR Protocol 2'* (*Quantitative Structure Activity Relationships\Appendix 07*), with three internally and externally valid models found using the GALib[53] subjective selection method, as detailed in table 6.12. The test sets for each of these models contained ten molecules, making them excellent sample sets with which to test the predictive capabilities of the models.[41]

Table. 6.12 Statistics for the three internally and externally significant NF54 models.

Model	Data	Method	Training set statistics					Test set statistics			
			r^2	q^2	F	r_{BS}^2	q^2	r^2	$(r^2-r_0^2)/r^2$	k	$\lvert r^2 - r_0^2 \rvert$
6	NF54	GALib-MLR	0.90	0.69	20.70	0.59	0.75	0.88	0.03	1.00	0.11
7	NF54	GALib-MLR	0.83	0.53	18.23	0.70	0.68	0.80	0.01	1.01	0.12
8	NF54	GALib-MLR	0.85	0.59	18.54	0.61	0.64	0.86	0.05	1.01	0.17

Table 6.13 gives the full statistics for one of the models in the series.

Table. 6.13 Statistics for model 7.

Model 7

Subjective selection method	GALib
Training set size	35
Test set size	10
Number of descriptors	7
Machine learning method	MLR
Molecule/Descriptor Ratio	5

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.83	Yes
q^2	0.72	Yes
F-statistic	18.23	Yes (Table value 2.37)
Bootstrapping	0.70	Yes (Ideally > 0.53)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.68	Yes
r^2	0.80	Yes
$(r^2 - r_o^2)/r^2$	0.01	Yes
k	1.01	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	0.12	Yes

Though the models looked encouraging, the LMO cross-validation approach was used for additional validation. The LMO cross-validation method involves dividing the dataset into several groups, and then building models for each of these subsets and calculating the q^2 statistic. Through repetition, a mean q^2 can be found, which ideally should not be much lower than the r^2 value calculated for the entire dataset. Usually a high value of q^2 (greater than 0.5) is considered proof of a models predictive ability,[35] and is a useful addition to evaluate a models performance. It is also a more rigorous validation procedure than LOO cross-validation.[22] Y-scrambling was performed for both the r^2 and LMO-q^2 statistics. Y-scrambling is useful to validate the robustness of a QSAR model, identifying those models based on chance correlation.[22] Models are developed and assessed when the dependent variable is randomly modified by assigning each compound a different dependent variable from the true set of responses.[39] If the original model is significant then

there should be a significant difference in the quality of the original model, and the average of those found using Y-scrambling.[37]

These additional statistical parameters for models 6 to 8 are shown in table 6.14. They were calculated using R 2.9.0,[54] which is a command line program used for statistical computing. The LMO cross-validation statistic was found by dividing the training set into five groups and then calculating the q^2 value for the data not used to develop the model. This was performed 1000 times and an average LMO value taken. Y scrambling was performed for both the r^2 and LMO-q^2 parameters using 1000 iterations.

Table. 6.14 Additional statistics for the three internally and externally significant NF54 models.

Model	Data	Method	Training set statistics			
			r^2	Average r^2 (Y scrambling)	LMO q^2	Average LMO q^2 (Y scrambling)
6	NF54	GALib-MLR	0.90	0.30	0.59	0.036
7	NF54	GALib-MLR	0.83	0.21	0.59	0.034
8	NF54	GALib-MLR	0.85	0.23	0.52	0.032

All three models had acceptable statistics, with r^2 values above 0.7, and LMO-q^2 values above 0.5. When Y scrambling was performed, these statistics were significantly lower as the data had been jumbled and therefore predictive models could not be found. This suggests that these models were not down to chance, and are more easily explained in graphical form. The first two graphs in figure 6.8 represent the histograms of the r^2 and LMO-q^2 values from Y scrambling for model 7. The vertical line represents the actual value for that statistic for the model, with the histogram showing the distribution of that statistic across the 1000 scrambled models. In both cases, the bars of the histogram are much lower than the actual value, indicating that scrambling the data does indeed affect the models, and thus it is unlikely that model 7 is down to chance. The third graph shows the relationship between r^2 and LMO-q^2 across all 1000 models, with the green spot representing

model 7. This offers further support for the quality of model 7. These graphs were similar for models 6 and 8 also, further supporting their quality.

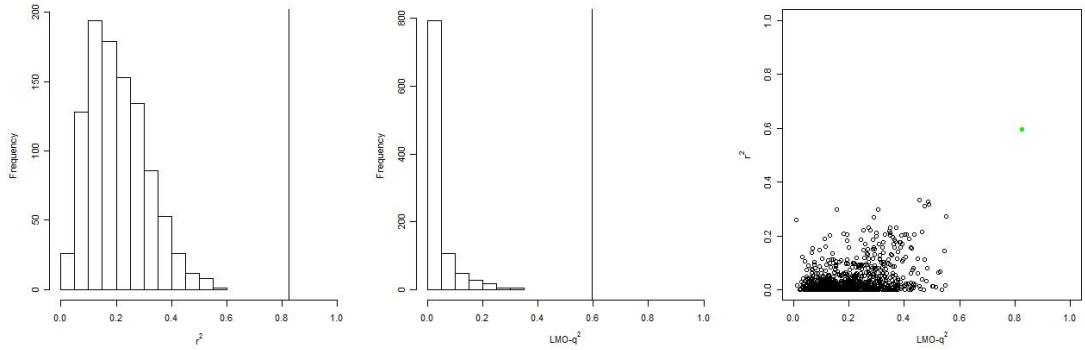


Fig. 6.8 Graphs illustrating the effect of Y scrambling on the r^2 and LMO-q^2 statistics for model 7.

The overall performance of the QSARs when only the 38 pre-selected descriptors were used produced much better models. Figure 6.9 illustrates the predicted versus actual activity values for both the training and test sets of models 7. As can be seen, unlike the previous models for the NF54 dataset, there is a clear linear relationship for both the training and test sets, indicative of the good predictive ability of the model. Importantly, it can also be seen that the data points are well distributed across the range of activities in both the training and test sets.

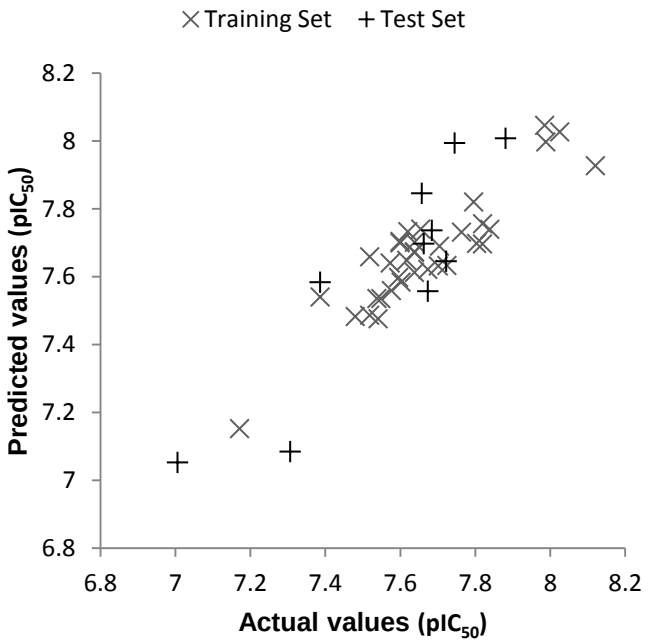


Fig. 6.9 Graph representing the predicted vs. actual activity values for model 7.

The most commonly occurring descriptors across the three significant models were identified as those which best described the NF54 activity. Only those descriptors which had a *t*-statistic greater than 2 were considered meaningful, with figure 6.10 illustrating the frequency of the descriptors across the models.

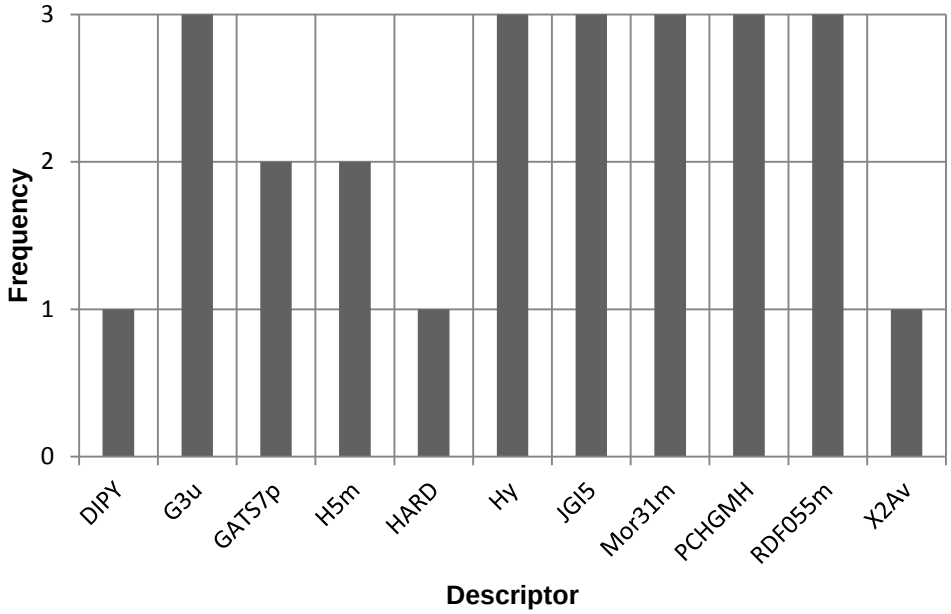


Fig. 6.10 Frequency of descriptors across the three significant NF54 models with a *t*-statistic greater than 2.

Six descriptors were found to be present in all three models (G3u; Hy; JGI5; Mor31m; PCHGMH; RDF055m) and were deemed to be of most interest. However, models were first developed for the K1 dataset to enable a comparison of the most important descriptors in modelling the activity of both strains. From this, conclusions could be drawn as to the chemical properties that effect the prediction of activity.

6.1.4.3 K1 Dataset

The same approach was therefore used to develop models for the K1 dataset when the 3D descriptors were included. As with the NF54 dataset, when all 1630 descriptors were first used to develop models for K1 via *'QSAR Protocol 2'*, this

yielded no externally valid models (*Quantitative Structure Activity Relationships\Appendix 08*). The descriptors were therefore split into two subsets (ADMEWORKS Modelbuilder[18] and DRAGON 3.0[19] descriptors) and *'QSAR Protocol 1'* was used to build models for each. Several internally significant models were identified (*Quantitative Structure Activity Relationships\Appendix 09*), with the descriptors from models that had a t-statistic greater than 2 considered to be those which best described the activity of the K1 strain. This gave 33 descriptors which were used to develop further models via *'QSAR Protocol 2'* (*Quantitative Structure Activity Relationships\Appendix 10*). Four internally and externally valid models were found using the GA and stepwise regression subjective selection methods, as shown by table 6.15.

Table. 6.15 Statistics for the four internally and externally significant K1 models.

Model	Data	Method	Training set statistics				Test set statistics				
			r^2	q^2	F	r_{BS}^2	q^2	r^2	$(r^2-r_0^2)/r^2$	k	$\lvert r^2 - r_0^2 \rvert$
9	K1	GA-MLR	0.91	0.83	35.09	0.78	0.72	0.75	0.019	1.00	0.011
10	K1	GA-MLR	0.91	0.83	33.48	0.72	0.57	0.68	0.0027	1.00	0.11
11	K1	Stepwise	0.68	0.56	22.45	0.53	0.88	0.89	0.038	1.00	0.031
12	K1	Stepwise	0.74	0.62	25.06	0.56	0.59	0.68	0.018	1.00	0.051

The test sets for models 9 to 12 contained a varying number of molecules, with one of the best performing models built using only 31 compounds, and externally validated on the other 14. The statistics for this model (model 12) are shown in table 6.16.

Table. 6.16 Statistics for model 12.

Model 12		
Subjective selection method	Stepwise regression	
Training set size	31	
Test set size	14	
Number of descriptors	3	
Machine learning method	MLR	
Molecule/Descriptor Ratio	10.333	
Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.74	Yes
q^2	0.62	Yes
F-statistic	25.06	Yes (Table value 2.96)
Bootstrapping	0.56	Yes (Ideally > 0.44)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.59	Yes
r^2	0.68	Yes
$(r^2 - r_o^2)/r^2$	0.018	Yes
k	1.00	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	0.051	Yes

As with the NF54 models, additional validation of models 9 to 12 was sought using LMO cross-validation, and Y scrambling of the r^2 and LMO-q^2 statistics (1000 iterations). The results of this analysis can be seen in table 6.17.

Table. 6.17 Additional statistics for the four internally and externally significant K1 models.

Model	**Data**	**Method**	**Training set statistics**			
			r^2	Average r^2 (Y scrambling)	LMO q^2	Average LMO q^2 (Y scrambling)
9	K1	GA-MLR	0.91	0.28	0.82	0.042
10	K1	GA-MLR	0.91	0.23	0.82	0.053
11	K1	Stepwise	0.68	0.085	0.56	0.046
12	K1	Stepwise	0.74	0.10	0.62	0.057

These statistics offered further validation of the predictive nature of the four models, suggesting that the QSARs were meaningful and not just down to chance. Figure 6.11 illustrates the results of Y scrambling for the best model (model 12), with the bars of the histograms being much lower than the actual model values for both r^2 and LMO-q^2. Models 9, 10 and 11 reported similar graphs.

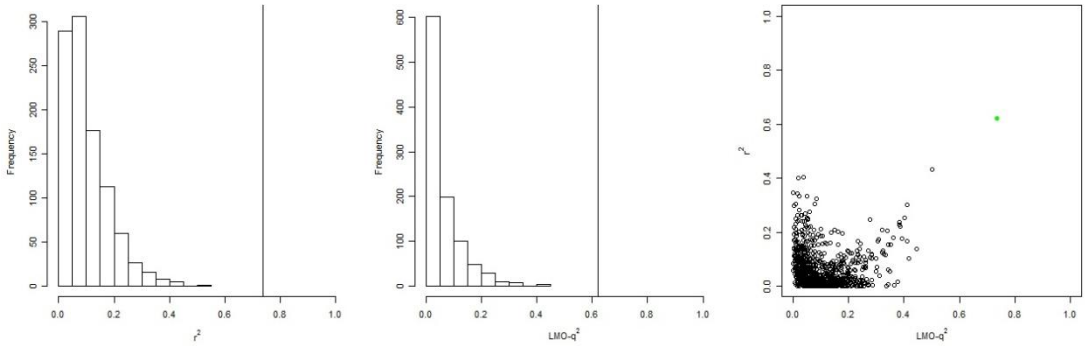


Fig. 6.11 Graphs illustrating the effect of Y scrambling on the r^2 and LMO-q^2 statistics for model 12.

With the models sufficiently validated it could be concluded that successful QSARs had been developed to model the K1 data, with figure 6.12 clearly illustrating the linear relationship observed between the training and test sets for model 12.

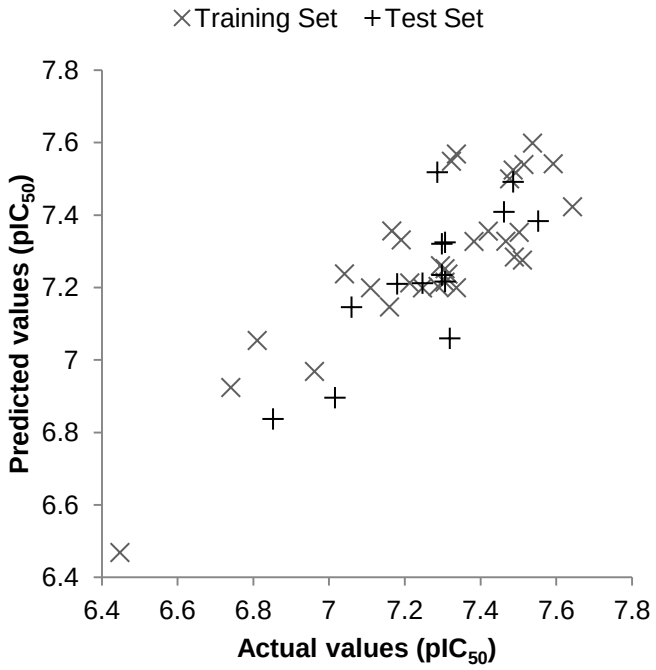


Fig. 6.12 Graph representing the predicted vs. actual activity values for model 12.

The commonly occurring descriptors across the four valid K1 models which had a t-statistic greater than 2 were identified, as was done with the NF54 models. The frequency of these descriptors is shown in figure 6.13. Only the Mor31e descriptor appeared in all four of the models, but several others did appear in half of them (i.e. DIPY, HARD, Rings count).

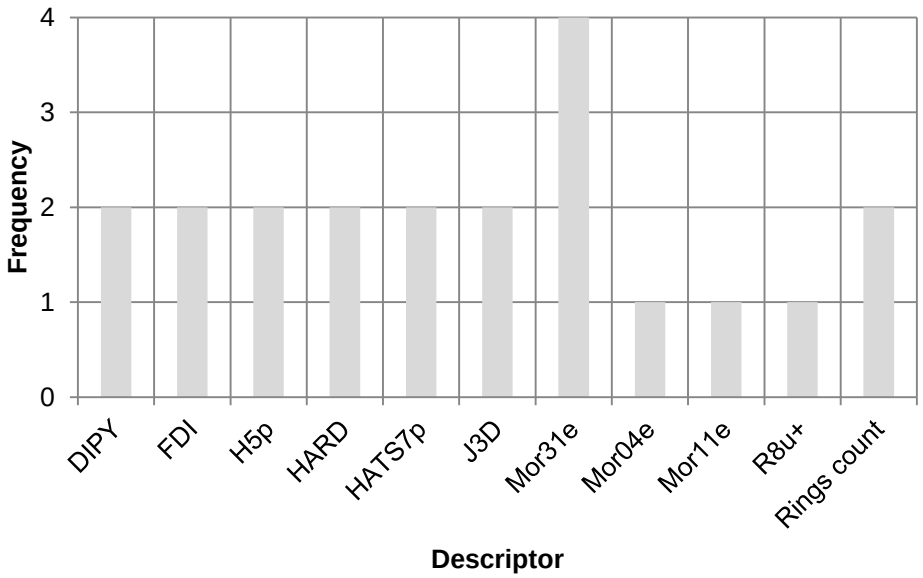


Fig. 6.13 Frequency of descriptors across the four significant K1 models with a *t*-statistic greater than 2.

Figure 6.14 shows the frequency of the molecular descriptors for both of the strains. DIPY and HARD were found to be the only descriptors common to models from both strains, whilst many others were unique to a particular strain. These descriptors were hoped to provide useful insight with regard to the nature of resistance between the two strains.

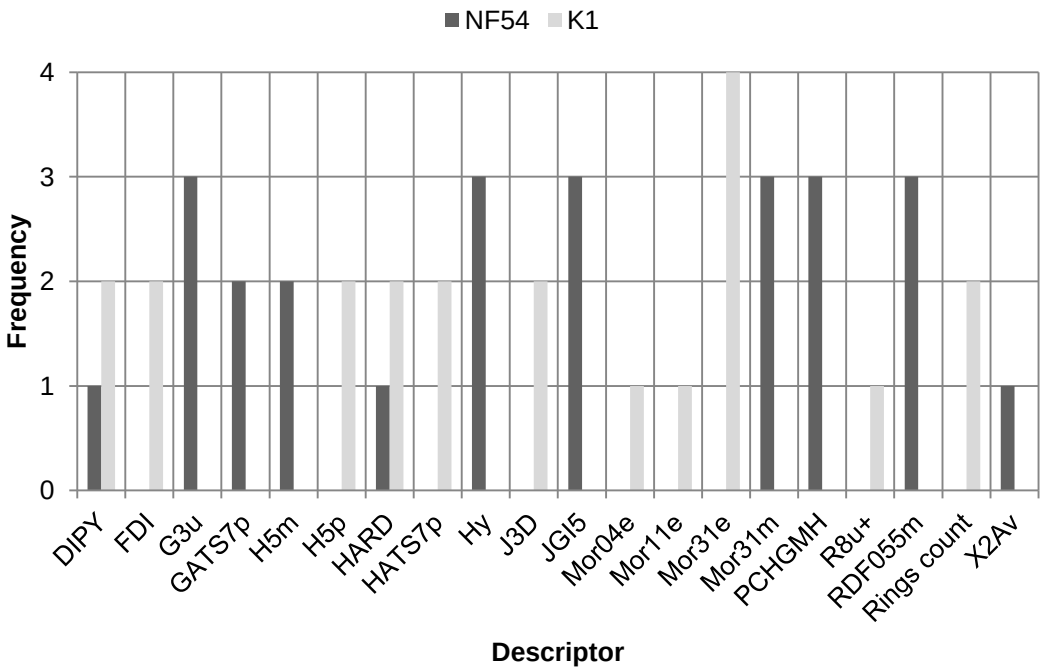


Fig. 6.14 Comparison of descriptor frequencies for the NF54 and K1 models.

Table 6.18 gives a brief definition of each of the descriptors in figure 6.14.

Table. 6.18 Descriptor abbreviations.

Abbreviation	Description
DIPY	Dipole Moment Y
FDI	folding degree index
G3u	3st component symmetry directional WHIM index / unweighted
GATS7p	Geary autocorrelation - lag 7 / weighted by atomic polarizabilities
H5m	H autocorrelation of lag 5 / weighted by atomic masses
H5p	H autocorrelation of lag 5 / weighted by atomic polarizabilities
HARD	Hardness
HATS7p	leverage-weighted autocorrelation of lag 7 / weighted by atomic polarizabilities
Hy	hydrophilic factor
J3D	3D-Balaban index
JGI5	mean topological charge index of order5
Mor04e	3D-MoRSE - signal 04 / weighted by atomic Sanderson electronegativities
Mor11e	3D-MoRSE - signal 11 / weighted by atomic Sanderson electronegativities
Mor31e	3D-MoRSE - signal 31 / weighted by atomic Sanderson electronegativities
Mor31m	3D-MoRSE - signal 31 / weighted by atomic masses
PCHGMH	Mean partial charge on H atoms
R8u+	R maximal autocorrelation of lag 8 / unweighted
RDF055m	Radical Distribution Function - 5.5 / weighted by atomic masses
Rings count	Number of rings
X2Av	average valence connectivity index chi-2

6.1.5 Combinatorial MLR Calculations

To gather further support of the descriptors detailed in figure 6.14, work was performed using UNIX codes to run MLR calculations that could account for any possible combination of descriptors, up to a maximum of 23.[55] A set of criteria were set such that only the most useful information would be recorded, with separate models built for both strains using the same 20 molecular descriptors for each. These 20 descriptors were those described in figure 6.14, as these had previously proven to be significant across both datasets. Only models with an r^2 value greater than 0.7, and between 1 and 9 descriptors were reported

For the NF54 dataset there were 41 models which fit these criteria, the results of which are shown in table 6.19. The frequency of the descriptors within these 41

models is expressed numerically and as a percentage, as well as the percentage population of the descriptor across all models.

Table. 6.19 Descriptor frequencies across the 41 models generated via combinatorial MLR calculations for the NF54 dataset.

Descriptor	Frequency (Number)	Frequency (Percentage)	Population of descriptor (Percentage)
Hy	41	100.00%	13.44%
JGI5	40	97.56%	13.11%
RDF055m	39	95.12%	12.79%
Mor31m	37	90.24%	12.13%
H5m	34	82.93%	11.15%
PCHGMH	28	68.29%	9.18%
HARD	25	60.98%	8.20%
DIPY	17	41.46%	5.57%
H5p	9	21.95%	2.95%
FDI	7	17.07%	2.30%
GATS7p	7	17.07%	2.30%
Mor31e	6	14.63%	1.97%
HATS7p	4	9.76%	1.31%
Mor04e	3	7.32%	0.98%
J3D	2	4.88%	0.66%
Rings count	2	4.88%	0.66%
G3u	2	4.88%	0.66%
R8u+	1	2.44%	0.33%
Mor11e	1	2.44%	0.33%

The descriptor Hy was present in all 41 models, whilst JGI5 was found in all but one. This is in strong accordance with the results in figure 6.14. Additionally, the work showed that the descriptor HARD was of moderate importance within many of the models.

The same combinatorial MLR analysis was performed for the K1 dataset, with 61 models found to have fewer than 23 descriptors and an r^2 value greater than 0.7. These results are shown in table 6.20.

Table. 6.20 Descriptor frequencies across the 61 models generated via combinatorial MLR calculations for the K1 dataset.

Descriptor	Frequency (Number)	Frequency (Percentage)	Population of descriptor (Percentage)
Mor31e	61	100.00%	15.48%
J3D	53	86.89%	13.45%
HARD	53	86.89%	13.45%
DIPY	50	81.97%	12.69%
HATS7p	29	47.54%	7.36%
JGI5	25	40.98%	6.35%
Mor04e	23	37.70%	5.84%
H5p	21	34.43%	5.33%
PCHGMH	17	27.87%	4.31%
X2Av	16	26.23%	4.06%
Rings count	15	24.59%	3.81%
FDI	11	18.03%	2.79%
H5m	9	14.75%	2.28%
Hy	6	9.84%	1.52%
Mor31m	3	4.92%	0.76%
R8u+	1	1.64%	0.25%
G3u	1	1.64%	0.25%

The Mor31e descriptor was present in all models, with HARD also found to be particularly prevalent. Once again, this is in line with GA-MLR analysis (fig. 6.14). Given that HARD was present in models for both the NF54 and K1 datasets; this suggests that the hardness of the molecules plays a key role in their antimalarial activity against both strains. Thus far only QSAR models developed using MLR have been discussed. However, it can be useful to use a consensus of different approaches,[56, 57] to identify commonality between the results which will hopefully offer further support for conclusions drawn from descriptor analysis.

6.1.6 Partial Least Squares

Partial least squares (PLS) regression is a useful technique when the number of independent variables is comparable to, or much greater than, the number of compounds. It can lead to stable and highly predictive models, even when there is a large degree of correlation between the descriptors.[58] It is a statistical method which is a combination of MLR and principle components regression (PCR),[59] and is used

to explain the variance across the descriptors, whilst attempting to obtain a good correlation between the dependent and independent variables. In PCR, the principle components are used as variables in an MLR type equation. This often leads to a concise QSAR equation as shown by equation 6.15, with the principle components selected based on their ability to explain the variance of the independent variables.

$$y = a_1 PC_1 + a_2 PC_2 + a_3 PC_3 + \cdots$$

Eq. 6.15 Linear equation using principle components.

PLS differs from PCR in that the components are calculated to explain the variation in both the independent and dependent variable, with the purpose being to find a small number of relevant factors that are predictive of the dependent variable, utilising the independent variables most efficiently.[60] PLS expresses the dependent variable in terms of quantities termed latent variables (t_i), as oppose to principle components seen in PCR. This gives an MLR equation of the form shown in equation 6.16.[1]

$$y = a_1 t_1 + a_2 t_2 + a_3 t_3 + \cdots + a_n t_n$$

Eq. 6.16 PLS regression equation.

The latent variables (t_i) are themselves linear combinations of the independent variables (x_i) as shown by equations 6.17.[1]

$$t_1 = b_{11} x_1 + b_{12} x_2 + \cdots + b_{1p} x_p$$
$$t_2 = b_{21} x_1 + b_{22} x_2 + \cdots + b_{2p} x_p$$
$$t_i = b_{i1} x_1 + b_{i2} x_2 + \cdots + b_{ip} x_p$$

Eq. 6.17 Latent variable equations.

The number of latent variables that can be generated is the smaller of the number of descriptors, or the number of observations. The latent variables are orthogonal to each other, and when calculated, PLS takes into account not only the variance in the

x variables, but also how this corresponds to the values of the dependent variable. The first latent variable (t_1) is a linear combination of the x values which give a good explanation of the variance in the x-space. However, it is also defined so that when multiplied by its corresponding coefficient (a_1), it provides a good approximation of the variation in the dependent variable. The second latent variable can then be determined and is orthogonal to the first, enabling more of the variation in the x dimension to be described, and for the prediction of the dependent variable to be further improved.

Analysis of PLS models can provide information about how the variables in a dataset can be combined to form a quantitative relationship between x and y. The weights of the various components provide understanding of which x variables are important, and which are uninformative,[22] with PLS models subject to the same level of scrutiny as those from MLR (i.e. internal and external validation). A major advantage of PLS is that unlike MLR, having a small number of descriptors is not a prerequisite for a meaningful model, in fact, sometimes a larger number of descriptors can create a better predictive model.[58, 61]

6.1.6.1 PLS Models

A series of QSAR models were developed for both the NF54 and K1 datasets (*Quantitative Structure Activity Relationships\Appendix 11* and *12* respectively) using *'QSAR Protocol 3'* as described in the Experimental Chapter, based on the previously selected descriptor subsets for each strain which had shown to be most promising (38 for NF54, 33 for K1).

6.1.6.1.1 NF54 Dataset

Table 6.21 gives the statistics for the best performing NF54 model, in which all but the internal bootstrapping statistic indicated a good model.

Table. 6.21 Statistics for model 13.

Model 13

Training set size	37	
Test set size	8	
Number of components	7	
Machine learning method	PLS	
Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.97	Yes
q^2	0.60	Yes
F-statistic	4.04×10^9	Yes (Table value 2.35)
Bootstrapping	0.45	No (Ideally > 0.67)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.72	Yes
r^2	0.84	Yes
$(r^2 - r_o^2)/r^2$	0.007	Yes
k	1.01	Yes
$\lvert r - r_o'^2 \rvert$	0.07	Yes

Figure 6.15 illustrates the predicted versus actual activity values for this model, for both the training and test sets. As can be seen, strong linear relationships were observed for both.

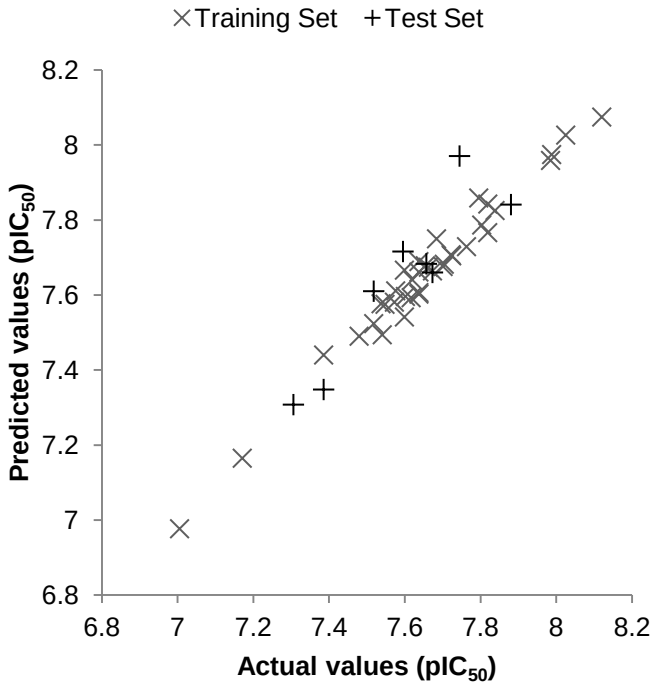


Fig. 6.15 Graph representing the predicted vs. actual activity values for model 13.

6.1.6.1.2 K1 Dataset

Table 6.22 gives the statistics for the best performing K1 model which passed all internal and external validation testing.

Table. 6.22 Statistics for model 14.

Model 14

Training set size	31
Test set size	14
Number of components	4
Machine learning method	PLS

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.98	Yes
q^2	0.75	Yes
F-statistic	5.72×10^9	Yes (Table value 2.74)
Bootstrapping	0.77	Yes (Ideally > 0.68)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.80	Yes
r^2	0.85	Yes
$(r^2 - r_o^2)/r^2$	0.01	Yes
k	1.01	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	0.01	Yes

Figure 6.16 illustrates the predicted versus actual values for this model, for both the training and test sets. As with the model for the NF54 dataset, strong linear relationships were observed.

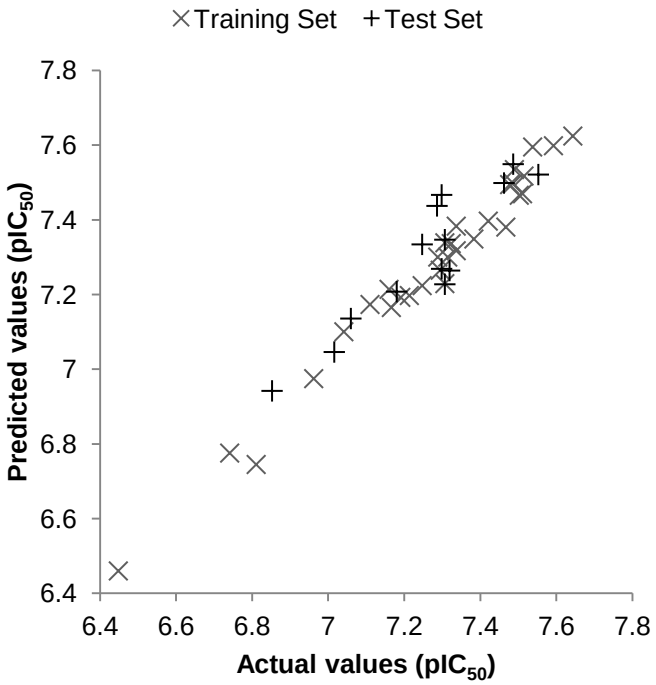


Fig. 6.16 Graph representing the predicted vs. actual activity values for model 14.

Generally, models developed using PLS produced excellent results, with the K1 models performing slightly better than those from NF54 analysis. The K1 models also required fewer components and explained a larger amount of data. As PLS uses components not descriptors, it was necessary to look at the weights of the descriptors within these components to assess their respective contributions. The descriptors with the largest weights added more to a particular component, and were therefore considered most useful in describing activity (*Quantitative Structure Activity Relationships\Appendix 13*).

6.1.7 Descriptor Frequencies

By comparing the results of the MLR and PLS models a set of potentially useful descriptors were identified. Some of these were common to both strains whilst others were unique to each. Descriptors present in both strains were thought to be important in order to explain why these compounds were initially active, whilst those present in only one strain could highlight possible differences between the two. It was thought that this information may be useful in understanding the resistance mechanism observed towards these 4-aminoquinoline compounds.

The two most important descriptors to model the NF54 dataset were Hy and JGI5, as they were the most frequently occurring descriptors for both the MLR and PLS models. Mor31e was found to be the most important/frequently occurring descriptor for modelling the K1 data, with the HARD descriptor fairly common across models from both strains of the parasite.

6.1.8 Descriptor Interpretations

Whilst it is not always essential to extract some sort of chemical meaning from a model, as it is its predictive ability which is most important, it can be useful to interpret the molecular descriptors and is a recommended practice.[22] Some descriptors such as log *P* or molecular weight are easy to interpret, whilst others can be more difficult. For several of the descriptors highlighted to be of interest from the study, it was difficult to infer some sort of easily understandable chemical meaning, however, what follows is an attempt to do so.

6.1.8.1 Hy

The Hy descriptor represents the hydrophilic factor, or hydrophilicity index. It was introduced by Todeschini and Gramatica[62] as a measure for the hydrophilic properties of a compound, and thus is (negatively) related to its hydrophobic properties.[63] Hy is a simple empirical index related to the hydrophilicity of a compound based on its substituent's, and can be calculated using equation 6.18.[62] In the equation, N_{Hy} represents the number of hydrophilic groups (-OH, -SH, -NH), N_C the number of carbon atoms and A the number of atoms (excluding hydrogen).

$$Hy = \frac{\left(1 + N_{Hy}\right).\log_2\left(1 + N_{Hy}\right) + N_C.\left(\frac{1}{A}.\log_2\frac{1}{A}\right) + \sqrt{\frac{N_{Hy}}{A^2}}}{\log_2(1 + A)}$$

Eq. 6.18 Hydrophilic factor equation.

Statistically Hy was identified as the most important molecular descriptor across the NF54 models, with the observed positive coefficient value of Hy suggesting that increasingly hydrophilic compounds have improved activity.[64] None of the K1 models contain the Hy descriptor, so this may have some reflection towards the mechanism of resistance in the K1 parasite. It would therefore be plausible to suggest, that given the hydrophilic factor of a molecule offers no indication as to how it will behave in the resistant strain, perhaps there has been some sort of hydrophobic change within the parasite, resulting in the acquisition of resistance. Either way, there has been some sort of change which has effected how hydrophilic molecules behave at the active site, because whilst the Hy descriptor was extremely important in exploring NF54 activity, it has no bearing on the models for K1.

6.1.8.2 JGI5

The JGI5 descriptor represents the mean Galvez topological charge index of order 5.[65] Topological charge indices were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfers in a given molecule.[66-68] Since many important physical, chemical, and biological properties are related to the charge distribution, the introduction of a topological index which could characterise this essential property was necessary.

Given the negative coefficient value of JGI5 across the NF54 models, it is possible that molecules which have a lower global charge transfer have better activity against the NF54 strain of the parasite, with only a moderate influence in the K1 strain (table 6.20).

6.1.8.3 Mor31e

Mor31e is a 3D-MoRSE descriptor which encodes for signal 31 weighted by atomic Sanderson electronegativities.[65] 3D-MoRSE descriptors are 3D-molecular representations of structures based on electron diffraction. They are based on the idea of obtaining information from the 3D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves.[69]

Very little information is available with regard to the chemical applications of the Mor31e descriptor, but it was found to be the most essential descriptor to explain and predict the activity of molecules against the K1 strain of the parasite. The negative coefficient values of Mor31e across the models indicates that it has a negative effect on the activity of compounds against K1, and thus compounds with a smaller Mor31e value may give improved activity against K1. A potential avenue of

exploration would be to consider calculating the Mor31e descriptor for a series of compounds and select those which report the lowest values, or to simply apply the full QSAR equation described in table 6.16 (model 12).

6.1.8.4 HARD

The HARD descriptor describes the hardness of a compound, and is calculated using MOPAC[70] (Molecular Orbital PACkage). MOPAC is a widely used semi-empirical quantum mechanical software package with a wide range of functionality.[71] HARD measures the resistance to change of electron distribution in a collection of nuclei and electrons.[72] It comes from the HSAB concept, which is an acronym for 'hard and soft acids and bases', also known as the Pearson acid base concept.[73] This concept is widely used in chemistry for explaining the stability of compounds, reaction mechanisms and pathways, as well as to assign the terms 'hard' or 'soft', and 'acid' or 'base' to chemical species. 'Hard' applies to species which are small, have high charge states and are weakly polarisable. 'Soft' applies to species which are big, have low charge density and are strongly polarisable.

The negative coefficient values of the HARD descriptor across the NF54 and K1 models suggests that 'hard' molecules have a negative impact on activity against the parasite, with 'soft' molecules i.e. those which are big and have low charge density, having much higher activity.

6.1.9 Descriptors and Chloroquine Resistance

As the Hy descriptor was the most important descriptor in describing the NF54 activity, and was absent from all the K1 models (bar 9.84% of the models calculated through combinatorial MLR analysis), it was therefore of little use in explaining the

activity of molecules toward the CQR strain of the parasite. However, through closer inspection of the descriptor it was possible to comment on the relationship between Hy and the biological activity of the 4-aminoquinoline compounds. As hydrophilic compounds are typically polar and capable of forming hydrogen bonds, they are more likely to become charged under the right conditions. This lends itself well to the theory as to how CQ and other aminoquinoline compounds become trapped in the DV. These hydrophilic compounds can become doubly protonated and membrane impermeable, thus leading to their accumulation in the acidic food vacuole. The fact that none of the K1 models make use of the Hy descriptor suggests that whether the compounds are easily protonated or not, has no bearing on their activity against the CQR strain, suggesting other factors must be at play. The molecules can easily be removed from the DV because of the mutation replacing the positive lysine with the neutral threonine. These results may therefore act to support the findings of others with regard to explaining CQ drug resistance.[74]

6.1.10 *k*-Nearest Neighbour

Though MLR and PLS were proven to give comprehensive results, additional work was carried out to develop models using the *k*-Nearest Neighbour (*k*NN) machine learning method, which is conceptually one of the simplest machine learning algorithms. Unlike regression methods which make the assumption that relationships are linearly related, *k*NN provides a non-linear method for deriving QSAR models. The *k*NN technique is a simple approach to pattern recognition problems,[75] with an unknown pattern being classified according to the majority of the class memberships of its *k*-nearest neighbours in the training set. The nearness is measured by an appropriate distance metric (i.e. a molecular similarity measure as

applied to the classification of molecular structures). The standard *k*NN method is implemented simply as follows:[76]

1. Calculate distances between an unknown object (u) and all the objects in the training set
2. Select k objects from the training set most similar to object u, according to the calculated distances (k is usually an odd number)
3. Classify object u with the group to which a majority of the k objects belong

An optimal k value is selected by the optimisation through the classification of a test set of samples, or via leave-N-out cross-validation. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k*-nearest neighbours. k is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its nearest neighbour. In binary (two class) classification problems, it is helpful to choose k to be an odd number, as this avoids tied votes.

The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbours. It can be useful to weight the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the distant ones.

The neighbours are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbours, the objects are represented by position vectors in a multidimensional feature space. It is usual to use Euclidean distance,

though other distance measures, such as the Manhattan distance could in principle be used instead. The *k*-nearest neighbour algorithm is however sensitive to the local structure of the data.

The *k*NN-QSAR method combines the *k*NN classification principle with the variable selection procedure. For each predefined number of variables (nvar) it seeks to optimize the following using stochastic sampling and simulated annealing as an optimisation tool.

1. The number of nearest neighbours (*k*) used to estimate the activity of each compound
2. Selection of variables from the original pool of all molecular descriptors that are used to calculate similarities between compounds

The upper limit of *k* is the total number of compounds in the data set, however, the best value has been found empirically to lie between 1 and 5.[75] The philosophy of *k*NN is straightforward. Since structurally similar compounds should have similar biological activities,[77] then the activity of a compound can be predicted (or estimated) simply as the average of the activities of similar compounds.

6.1.10.1 *k*NN Models

'QSAR Protocol 4' as described in the Experimental Chapter outlines the method which was used in order to develop a series of models for the NF54 and K1 datasets, using the preselected descriptor subsets (38 and 33 respectively). The main goal of kNN-QSAR was to see whether a non-linear method could identify any meaningful models. The procedure involved using codes which generate large numbers of models for varying training and test set splits, optimising *k* for each one. Those

which pass the required statistical parameters are recorded, and the subsequent models undergo randomization, with successful models grouped according to their significance.

The randomization step produced a number of models for each of the strains. The majority of the models for K1 were significant to 0.005%, whereas most of the models for NF54 were significant to only 1%. The lower the percentage with which the models are significant, then the stronger the performance of the models, and the less likely it is that the relationships are down to chance. The models for K1 were therefore statistically better than those for NF54. This may be due to the characteristics of the datasets. NF54 represents the sensitive strain, so a much broader set of chemical diversity is accommodated at the site of action, exhibited by the fact that CQ is active. However, the activities of the molecules against the K1 strain are dependent on a much more stringent set of criteria, so perhaps it is easier for this machine learning method to build models for this dataset. This observation also holds true for the other machine learning methods employed, as generally the models for the K1 strain performed better than their NF54 counterparts.

Figure 6.17 represents the average of the predicted values across all models, versus the actual values for the NF54 strain at 1% significance, including their SD values (*Quantitative Structure Activity Relationships\Appendix 14*).

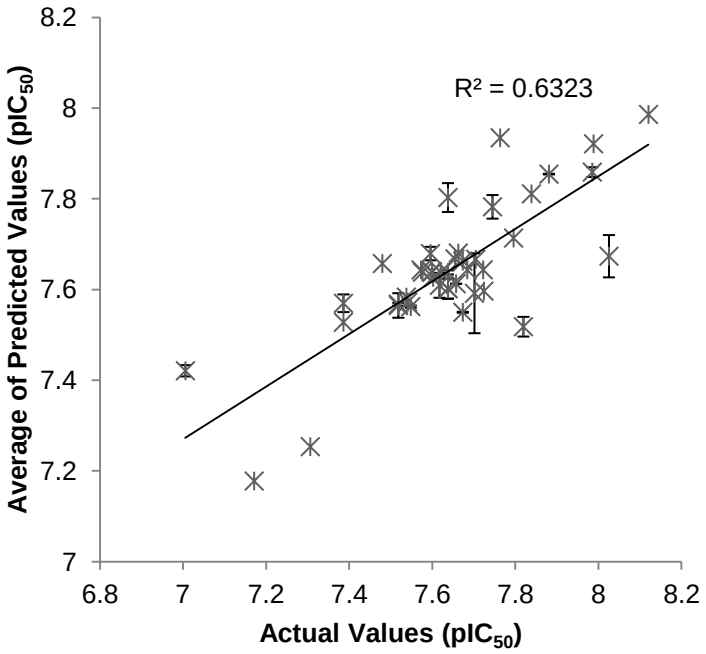


Fig. 6.17 Average of the predicted values across all models vs. the actual values for the four NF54 models significant to 1%.

Only four models were found at the 1% significance level, with some of the molecules only being represented once. This is why there appears to be very small error bars for some of the molecules, or none at all. The data had an r^2 value of 0.63, which though encouraging, is not as promising as the results from MLR and PLS analysis.

Figure 6.18 represents the average of the predicted values across all models, versus the actual values for the K1 strain at 0.005% significance, including their SD values (*Quantitative Structure Activity Relationships\Appendix 14*).

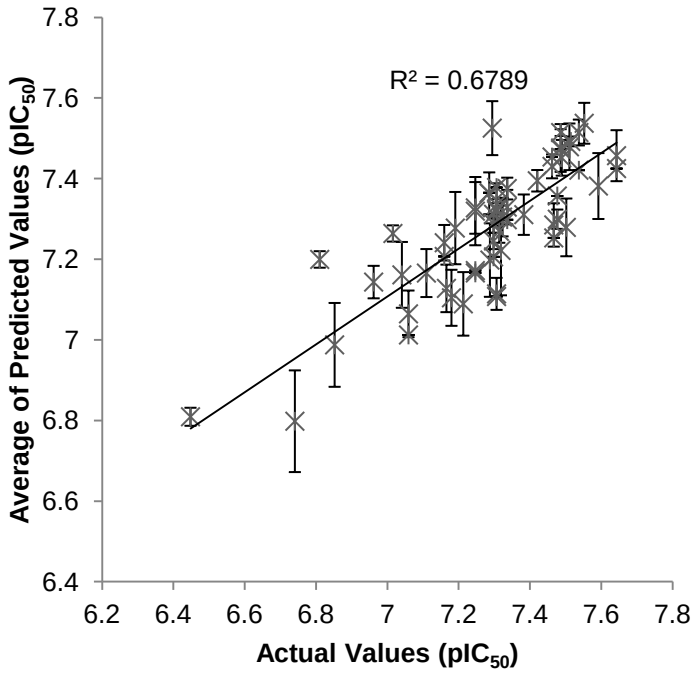


Fig. 6.18 Average of the predicted values across all models vs. the actual values for the eight K1 models significant to 0.005%.

The results are better than those for NF54, as there were eight models found to be significant at 0.005%, with the data having a r^2 value of 0.68. There is very little correlation/overlap between the descriptors used in the models for each strain, therefore the results are very limiting when it comes to extracting some sort of chemical meaning, as none stand out as more important than others. The frequency of the descriptors across the four NF54 models is represented by table 6.23. What is interesting to note is that none of the previously identified descriptors of interest (Hy and JGI5) were present in these models, nor was the HARD descriptor which was common to models for both strains. This observation does not raise too much concern however, as the descriptors in *k*NN analysis are used in a non-linear fashion, and as such are less easy to interpret, compared to those from linear methods (i.e. the descriptor coefficient values).

Table. 6.23 Frequency of descriptors across the four NF54 models.

Descriptors	Frequency of Descriptor Across the Models
DIP	4
DIPX	4
H5m	4
Mor20e	4
PCHGMH	4
R1e	4
RDF035e	4
X2Av	4
C-008	3
GATS4e	3
Mor24e	3
R3u+	3
RDF055m	3
G3u	2
H-046	2
H5p	2
GGI1	1
Mor31m	1

The frequency of the descriptors across the eight K1 models was calculated and is represented by table 6.24. The Mor31e descriptor which was common to most of the linear models was present in only three of the eight *k*NN models, whilst the HARD descriptor appeared in all of them. These models are therefore much more supportive of the earlier conclusions.

Table. 6.24 Frequency of descriptors across the eight K1 models.

Descriptors	Frequency of Descriptor Across the Models
FDI	8
HARD	8
HATS7p	8
Mor20e	8
DIPX	7
RDF065m	6
Mor04e	5
RDF030m	5
Mor31m	4
PCHGMHT	4
PW2	4
BELp3	3
ENEG	3
Mor31e	3
nCt	3
R2u	3
R8u+	3
SCOUNT(C-atom)	3
DIPY	2
G2v	2
GGI1	2
J3D	2
LOGP	2
RDF065m	2
RTu+	2
Mor11e	1

The main objective had however been achieved, which was to develop validated QSAR models for both strains using the *k*NN non-linear method. The interpretability of these models is limited however when compared to the linear methods, as the descriptors cannot be easily weighted within the models, and are therefore unrelated to the mode of action of the molecules.

This method created many training and test set splits to generate QSAR models, therefore in several of the statistically valid models, a number of the molecules were missing from the test set. The reason for this may be due to the domain of applicability of the molecules within that model. The *k*NN-QSARs were developed

by interpolating the activities of the *k* nearest neighbours for each compound to predict its activity.[75] This procedure derives a special applicability domain specific to each particular model, to avoid making predictions for compounds that differ substantially from the training set molecules.[37] The threshold D_T can be calculated from the training set models using equation 6.19. $\bar{y}$ is the average Euclidean distance between each compound and its *k* nearest neighbours (where *k* is the parameter optimized in the course of QSAR modelling), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. The default value of this parameter is set at 0.5, which formally places the allowed distance threshold at one-half of the standard deviation (assuming a Boltzmann distribution of distances between *k*-nearest neighbour compounds in the training set). Thus, if the distance of the external compound from at least one of its nearest neighbours in the training set exceeds this threshold, the prediction is considered unreliable, and the result is omitted.

$$D_T = \bar{y} + Z\sigma$$

Eq. 6.19 Threshold D_T calculation.

6.1.11 Selective/Resistance Index

Two distinct activity datasets for the same molecules can be compared through consideration of their selective or resistance index. Calculating the ratio of activities between the NF54 and K1 strains (*Quantitative Structure Activity Relationships\Appendix 16*) for each compound gave an indication as to the resistance of the molecules between the two strains. A higher value indicated a greater resistance for that particular molecule towards the K1 strain, with CQ therefore having the highest selective index value described in Patent 5596002.[12]

MLR models were first developed using *'QSAR Protocol 1'* in order to identify a suitable descriptor subset using all 45 compound in the training set and all 1630 0 to 3D molecular descriptors available. From the statistically significant models which were generated, 33 descriptors were selected as these each had t-statistics greater than 2. Further models were generated using this subset of descriptors according to *'QSAR Protocol 2'*. Whilst all of the models developed were internally valid (*Quantitative Structure Activity Relationships\Appendix 17*), externally they failed to meet all of the required parameters. Table 6.25 and figure 6.19 illustrate the results from one such model.

Table. 6.25 Statistics for model 15.

Model 15

Subjective selection method	Genetic algorithm
Training set size	38
Test set size	7
Number of descriptors	8
Machine learning method	MLR
Molecule/Descriptor Ratio	4.750

Internal Validation		
Parameter	**Value**	**Valid?**
r^2	0.85	Yes
q^2	0.70	Yes
F-statistic	20.59	Yes (Table value 2.28)
Bootstrapping	0.63	Yes (Ideally > 0.55)
External Validation		
Parameter	**Value**	**Valid?**
q^2	0.43	No
r^2	0.76	Yes
$(r^2 - r_o^2)/r^2$	0.01	Yes
k	1.01	Yes
$\lvert r_o^2 - r_o'^2 \rvert$	0.15	Yes

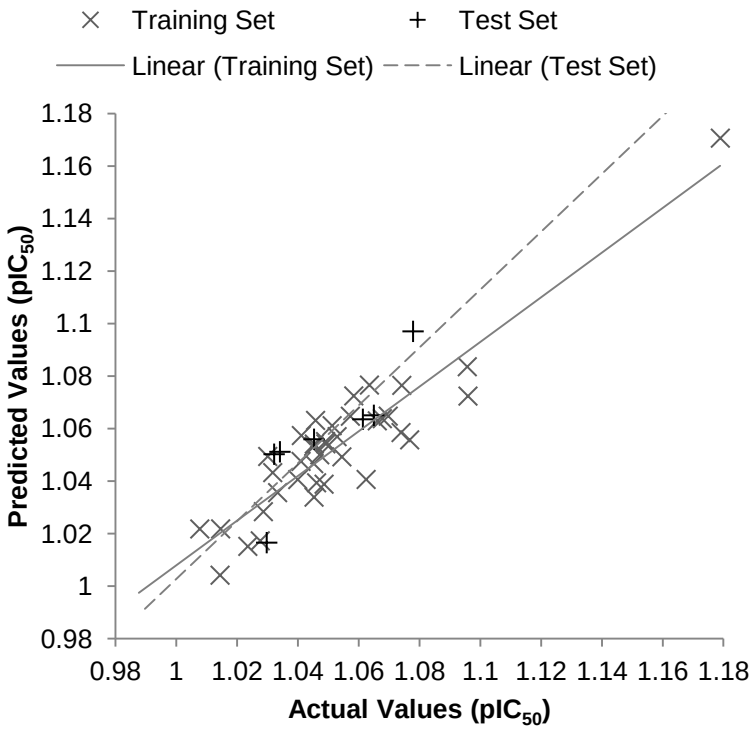


Fig. 6.19 Graph representing the predicted vs. actual selective index values for model 15.

In an ideal situation, the linear relationships between the training and test set would be as close to each other as possible. Here however they have different gradients, indicating the poor predictive ability of the model. This can also be seen by looking at the external q^2 statistic for model 15. The reason for this is most likely due to there being very little spread within the data. All of the selective index values are within a range of 0.18, with CQ appearing as an outlier in the data. This skews the results giving a cluster and a point, which becomes very difficult to accurately model. Additional calculations were performed with CQ removed, but this still failed to produce any externally valid models, supporting the conclusion that the range of the dataset was too narrow to model in this manner.

6.1.12 Summary of 4-Aminoquinoline QSAR Analysis

A series of statistically significant QSAR models were generated for the 45 4-aminoquinoline compounds described in Patent 5590002,[12] which had been tested

against both the NF54 and K1 strains of the malaria parasite. A host of machine learning methods were investigated, and the molecular descriptors contained within these valid models were interpreted with reference to the mechanistic mode of action of the 4-aminoquinolines. Considerations were also been put forward with regard to combating the observed resistance in the parasite towards this class of compound. Several significant models were identified which were shown to have strong predictive abilities for both strains of the parasite. These models can now be used to predict the antimalarial potential for new 4-aminoquinoline compounds, even before they have been synthesised. This will help to refine the molecular design loop, and may help to reduce the number of redundant compounds which are made, thus creating a more efficient development process, as well as saving both time and money.

6.2 Hepatitis C Thiazolides QSARs

An additional QSAR study was performed in order to develop QSAR models for a series of thiazolide compounds which were active as antiviral agents against the hepatitis C virus (HCV). This small study is now discussed.

6.2.1 Hepatitis C Virus

HCV is a single-stranded RNA virus of the *Flaviviridae* family that is responsible for hepatitis C in humans.[78] Hepatitis C is an infectious disease primarily affecting the liver. The infection is often asymptomatic, but chronic infection can lead to scarring of the liver and ultimately liver fibrosis, cirrhosis and cancer, which may appear up to 30 years later.[79-81] About 170 million people worldwide are chronically infected with HCV, which is transmitted principally through blood infection, and for which there is currently no vaccine available.[82] Many genotypes of HCV have been

recognised; with genotypes IA/IB accounting for about 70% of all cases, and genotypes IA/IB and IIA/IIB combined accounting for over 90%.[83] Evaluation of anti HCV agents therefore usually begins with genotypes IA/IB.

A number of therapeutic approaches to HCV treatment are possible.[78, 84] However, the combination of (pegylated) interferon α (IFN-α) and ribavirin (fig. 6.20) is still regarded as standard of care (SOC), even though a sustained virological response is only observed in 50-60% of patients, with genotype I infections being more difficult to treat.[85, 86] The mode of action of IFN-α-ribavirin is not wholly clear, but some kind of immunomodulatory effect does seem to be involved, and indeed a number of candidate anti-HCV therapies involve either new formulations of IFNs or other immune stimulants.[87]

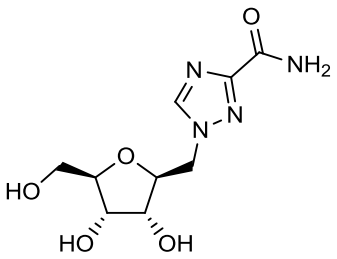


Fig. 6.20 Ribavirin.

Nitazoxanide (NTZ, fig. 6.21) is a broad spectrum thiazolide anti-infective which has been licensed in the US for the treatment of diarrhoea. It has also been shown to be effective against both DNA and RNA viruses, in particularly the hepatitis B virus (HBV) which causes hepatitis B, another infectious inflammatory disease of the liver.[88-90] It is currently thought that NTZ acts via modulation of host cell processes,[91, 92] and is an effective anti-HCV agent either alone or in combination with SOC treatment.[93-95] Several thiazolide analogues have also shown promising anti-viral activity, with *in vitro* activity values comparable to those of NTZ with good cell safety indexes.[96]

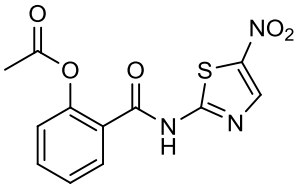
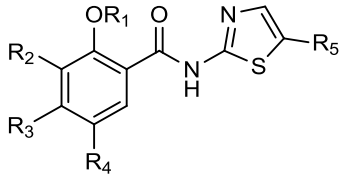


Fig. 6.21 Nitazoxanide.

It was hoped that by implementing QSAR methods using a number of thiazolide derivatives, good predictive models with which to model activity could be generated. A series of 27 compounds were considered, each of which had been tested against several assays for various HCV genotypes.[96] Though manual interpretation of the SAR led to some interesting observations, such as methylation of the benzene ring leads to a decrease in activity, QSAR analysis allows for more informative conclusions to be drawn, as well as models which could be used for future prediction. The 27 thiazolide derivatives could be split into two subsets according to their general structures. The first subset contained 15 5'-nitro- and 5'-halothiazolide derivatives of the general structure shown in table 6.26, which also reports the activity values for each compound. The primary assay was used to test against genotype IB, with CC_{50} and EC_{50} values reported for each compound in μM. The former represents the concentration at which 50% cell cytotoxicity was observed, and the later the half maximal effective concentration of the compound. Combined, these measure the cell safety and the *in vitro* activity of a molecule respectively. Compounds showing good efficacy in the primary assay were then subjected to a secondary assay of the same genotype, as well as a primary assay against genotype 1A.

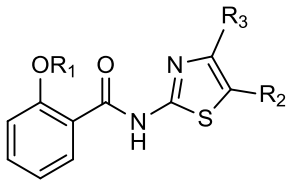
Table. 2.26 Activities of 5'-nitro and 5'-halothizolides against HCV replication.



Compound	R_1	R_2	R_3	R_4	R_5	Primary assay Genotype 1B		Secondary assay Genotype 1B		Secondary assay Genotype 1A	
						CC_{50} (μM)	EC_{50} (μM)	CC_{50} (μM)	EC_{50} (μM)	CC_{50} (μM)	EC_{50} (μM)
1	Ac	H	H	H	NO_2	38	0.21	35	0.25	49	0.33
2	H	H	H	H	NO_2	15	0.15	18	0.15	14	0.25
3	H	H	H	H	Cl	15	10				
4	Ac	H	H	H	Cl	4.3	0.23	4.9	0.31	5.7	0.4
5	H	Me	H	H	NO_2	5	0.36	11	0.35	12	0.39
6	H	H	H	Cl	NO_2	>100.0	2				
7	Ac	H	H	H	Br	15	3.8	12	4.3	21	3.3
8	H	H	H	H	Br	20	10	98	4.9	88	2.8
9	H	Me	H	H	Br	21	1.9	15	>10.0	25	>10.0
10	Ac	H	H	Me	Br	12	4.2	10	3	11	3
11	H	Cl	H	H	Br	14	5.2	10	>10.0	11	>10.0
12	Ac	Me	H	H	Cl	12	0.59	16	1.5	14	1
13	H	H	Me	H	Cl	30	>10.0				
14	H	H	H	Me	Cl	2.3	>20.0	20	10	16	10
15	H	H	H	H	F	24	>10.0			2.9	>10.0

Table 6.27 represents the second subset which consists of 12 4'- and 5'substituted thiazolides with the general structure shown, together with their activity.

Table. 6.27 Activities of 4'- and 5'-substituted thiazolides against HCV replication.



Compound	R_1	R_2	R_3	Primary assay Genotype 1B		Secondary assay Genotype 1B		Secondary assay Genotype 1A	
				CC_{50} (μM)	EC_{50} (μM)	CC_{50} (μM)	EC_{50} (μM)	CC_{50} (μM)	EC_{50} (μM)
16	Ac	CF_3	H	3.7	4.3	1.7	>10.0	1.4	>10.0
17	Ac	H	CF_3	2.8	>10.0				
18	H	Me	H	45	4.2				
19	H	NHAc	H	11	>10.0				
20	H	CO_2Et	H	14	>10.0				
21	H	H	Ph	26	3.5	14	>10.0	29	>10.0
22	H	H	SOMe	63	2				
23	H	H	SO_2Me	*85*	*0.5*				
24	H	SO_2Me	H	*43*	*1.5*				
25	H	Br	Me	15	>10.0				
26	H	Br	Ph	16	2.2	100	>10.0	62	>10.0
27	H	CN	H	15	3.7	15	>10.0	11	>10.0

Between the two subsets there were a total of 27 thiazolide derivatives. Similar to the analysis of the 4-aminoquinoline dataset, QSARs were developed using the following three steps:[13]

i. Data preparation

ii. Data analysis

iii. Model validation

6.2.2 Data Preparation

The 27 thiazolide derivatives were first constructed and energy minimised using the *'Energy Minimisation Protocol'*. 946 (0, 1 and 2D) molecular descriptors were then calculated for each of the molecules using DRAGON 6.0.[97] QSARs were developed for several of the assay datasets present in tables 6.26 and 6.27, however, in order to

assess the predictive ability of a model several considerations had to be put in place. A training set was required to contain a minimum of ten compounds,[40] whilst the external test set should have a minimum of five compounds.[98] As such, sufficient quantitative data existed for only four of the assays: primary assay genotype IB CC_{50} (26 data points); primary assay genotype IB EC_{50} (20 data points); secondary assay genotype IB CC_{50} (15 data points); secondary assay genotype IA CC_{50} (16 data points). These datasets allowed for sufficient splitting (large training and test sets), and also contained evenly distributed values with low µM or sub µM activities. The CC_{50} and EC_{50} values were converted into pCC_{50} and pEC_{50} values using negative log transformations in order to normalise the data and hopefully create better fits (*Quantitative Structure Activity Relationships\Appendix 18*).[14]

6.2.3 Model Development

Initially, models were constructed for the dataset using GA-MLR according to *'QSAR Protocol 1'* (*Quantitative Structure Activity Relationships\Appendix 19*), with models validated using the previously discussed criteria for internal validation (table 6.3). Table 6.28 gives the statistics for the best performing model for each of the four assays when using all molecules in the training set. As can be seen, a robust selection of models were found, each satisfying the specified internal validation criteria, and showing good correlations between measured and predicted activity values.

Table. 6.28 GA-MLR QSAR models and their performance statistics for the primary assay genotype IB CC_{50}, primary assay genotype IB EC_{50}, secondary assay genotype IB CC_{50}, secondary assay genotype IA CC_{50}.

	Primary assay genotype IB		**Secondary assay**	
	CC_{50}	**EC_{50}**	**Genotype IB CC_{50}**	**Genotype IA CC_{50}**
r^2	0.81	0.77	0.81	0.77
q^2	0.68	0.60	0.63	0.62
r^2_{BS}	0.60	0.50	0.51	0.48
F	21.73	12.29	15.37	13.38
Significant descriptors	BLI GATS8m P_VSA_MR_2 RDF030p	ATSC2p P_VSA_i_3 G2u R5s+	nX JGI9 RDF090m	SpMax_B(s) RDF040v G1v

One correlation is illustrated by figure 6.22, which shows the calculated against experimental activity values for the primary assay genotype 1B CC_{50}.

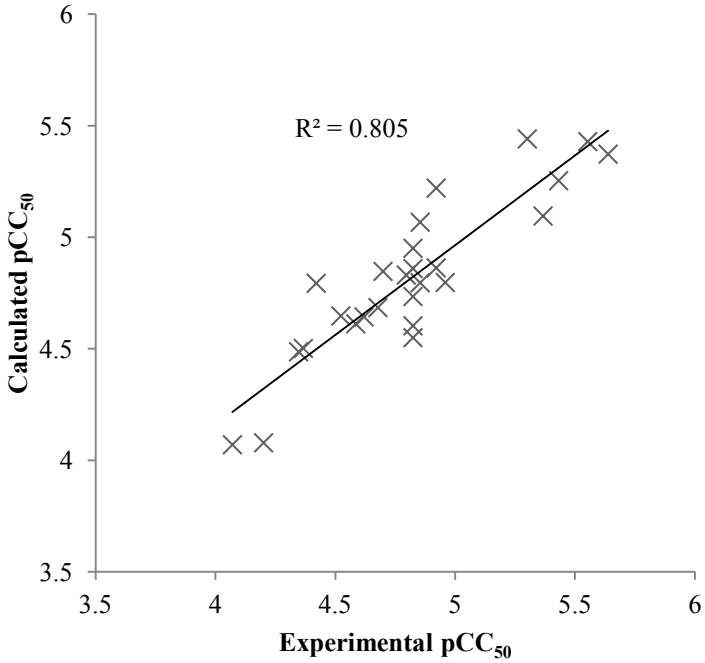


Fig. 6.22 GA-MLR regression model for the primary assay genotype IB CC_{50}.

The molecular descriptors used within each model were all significant as they observed t-statistic values greater than 2. However, with only one exception, the descriptors were challenging to interpret with respect to aiding further manual molecular design. The one interpretable descriptor was within the secondary assay IB CC_{50}, model and encoded the number of halogens in a compound (nX). For the other descriptors in the models a correlation with physically interpretable descriptors

was sought, with descriptors which had greater than or equal to 0.75 correlation identified. For the primary assay genotype IB EC_{50}, a high correlation was discovered between ATSC2p (negative correlation with respect to activity) and nH (number of hydrogen atoms) and N% (percentage of N atoms). Additionally R5s+ (negative correlation with respect to activity) was negatively correlated with nH. This presents apposing opinions with regard to the nH descriptor, and thus is of little use. Finally the SpMax_B(s) descriptor in the secondary assay genotype IA CC_{50} model observed negative correlation with activity, and was negatively correlated with nF (number of F atoms). Despite strong correlations, these descriptors are still of little use with reference to manual molecular design, and thus the mode of action for these compounds.

The models all perform very well internally, but as has been stated throughout this chapter, the only true evaluation of a models predictive ability comes through external validation.[38] When an internally significant model was developed it was to be applied to a test set and assessed using the external validation criteria specified in table 6.4. The four assay datasets which contained sufficient data points needed to be split in order for their predictive ability to be studied. To do this a number of splitting methods were considered. Initially the sphere-exclusion algorithm was used,[44] however, owing to the narrow range of activity values which spanned less than two orders of magnitude, this method had difficulties creating a suitable test set that represented the spread of the points suitably. For this reason the CADEX (Computer Adjunct Data Evaluator X) algorithm[99] was employed, with the test sets for each assay contained five molecules, satisfying the minimum number of molecules for the test set, yet leaving a suitably large training sets. Models were then developed using GA-MLR via *'QSAR Protocol 2'* (*Quantitative Structure*

Activity Relationships\Appendix 19). Despite all the models having good internal statistics, when they were applied to their test sets they fell apart, failing to pass the external validation criteria. Figure 6.23 represents the training and test set for one such failed model, in this case for the primary assay genotype 1B CC_{50} dataset. Though the r^2 value for the test set had an acceptable value ($r^2 = 0.66$), all the other statistics were poor. The reason for this may once again be due to the limited spread of the data points in the test set.

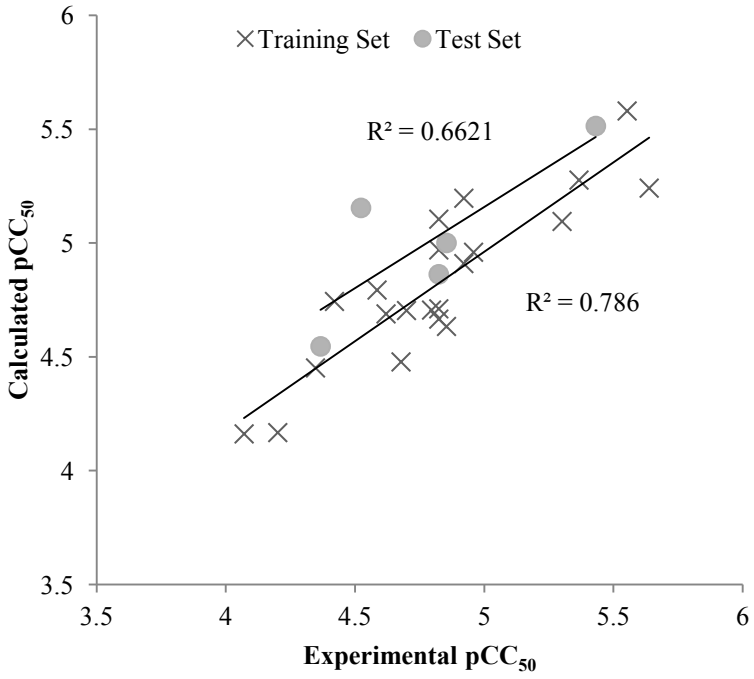


Fig. 6.23 Training and test set for the failed GA-MLR regression model of the primary assay genotype IB CC_{50}.

The final splitting method involved using activity binning. Each molecule within the various assays was assigned to a specific bin, with the central molecule from each bin selected for the test set. In cases where the bin contained an even number of molecules the highest median was selected, with this trend performed across all other even numbered bins. This was then repeated only with the lower median compounds selected, and QSARs developed for both test splitting patterns. However, even this

method of splitting failed to produce any meaningful models when developed using *'QSAR Protocol 2'*.

With linear QSAR methodology failing to produce any externally significant models despite a number of splitting methods being used, it was decided to explore the possibilities of a nonlinear method. SVM has been discussed briefly in Chapter I, and has proven to be a useful technique for addressing a wide range of classification and regression problems.[100, 101] SVM models were developed using a grid search approach through the workflow program KNIME,[102] as outlined by *'QSAR Protocol 5'* described in the Experimental Chapter. This grid search approach allowed for the parameters to be optimised in order to converge towards the best solutions. Splitting of the dataset was performed prior to model development, with the CADEX algorithm failing to produce any splits which went on to yield meaningful models (*Quantitative Structure Activity Relationships\Appendix 20*). Thus, activity binning was used and showed some success.

It was performed by creating eqully sized bins which represented a particular proportion of the spread in activity values. If five molecules were required in the test set, then the molecules were assigned to one of five unique bins. The central molecule from each bin was then chosen to represent the test set. As before, if an even number of molecules were in the bin then the highest median was selected, and QSARs built for this dataset, with the process repeated using the lowest median. Table 6.29 illustrates the external validation statistics for the best SVM QSAR models found using activity binning for each of the four assays (*Quantitative Structure Activity Relationships\Appendix 21*). A statistically signicant model, both internally and externally, was found for only the primary assay genotype 1B CC_{50}

dataset. This assay had the largest number of data points (26 compounds), and the subsequent model passed all internal and external criteria.

Table. 6.29 SVM QSAR models and their external validation performance statistics for the primary assay genotype IB CC_{50}, primary assay genotype IB EC_{50}, secondary assay genotype IB CC_{50}, secondary assay genotype IA CC_{50}.

	Primary assay genotype IB		**Secondary assay**	
	CC_{50}	**EC_{50}**	**Genotype IB CC_{50}**	**Genotype IA CC_{50}**
$\boldsymbol{r^2}$	0.8052	0.420	0.0109	0.5218
$\boldsymbol{q^2}$	0.561	0.396	-0.0329	0.1051
$\boldsymbol{\lvert r_0^2 - r_0'^2 \rvert}$	0.0001	0.0001	0	0
$\boldsymbol{(r^2 - r_0^2)/r^2}$	0.000103	0.000104	0	0
$\boldsymbol{k}$	1.0073	0.987	1.0096	1.0176

Figure 6.24 represents the plot of the training and test set for the primary assay genotype IB CC_{50} model. It can be seen that in this case activity binning selected a diverse set of compounds for the test set, which accurately represented the spread of activity points. Of the three assay datasets for which no externally valid models could be found, it is interesting to note that in all of these models, the training set had perfect correlation, with internal r^2 values of 1. This suggests that the data had been overtrained, and was therefore unsuitable for use in a predictive capacity. Also, the failed datasets had the smallest number of data points, with narrow activity windows, potentially explaining why poor models were found.

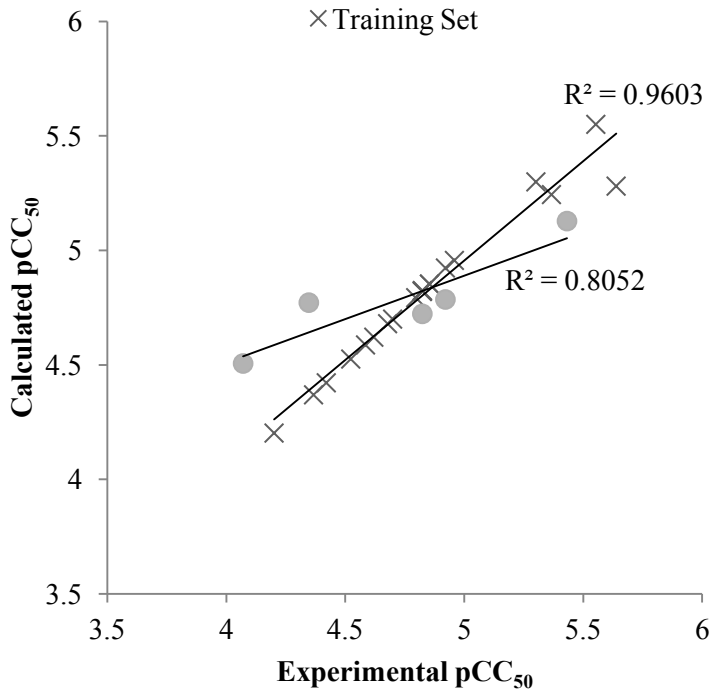


Fig. 6.24 Training and test set for the failed GA-MLR regression model of the primary assay genotype IB CC_{50}

6.2.4 Summary of Thiazolide QSAR Analysis

A significant model was developed for the primary assay genotype 1B CC_{50} dataset. It is hoped that this model will be useful for future compound selection and design, as it can successfully predict the cell safety indices of thiazolide derivatives. Though no meaningful models were found for the other assays, this is most likely due to the fewer number of data points available, so once additional data has been collected it is hoped that similarly predictive models will be developed.

Combined, the 4-aminoquinoline and thiazolide studies illustrate the power and potential of QSAR methods, as both have led to validated/predictive models which can be used to aid in drug design and safety respectively. In certain cases, QSAR may also be used to aid in biological interpretation, in order to gain insight into potential modes of activity and resistance.

Chapter VII is the final results chapter of this thesis and reports the design and synthesis of a novel series of pyrroloquinolone containing compounds, which were later shown to be active against malaria.

6.3 References

1. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
2. A. Sparatore, N. Basilico, S. Parapini, S. Romeo, F. Novelli, F. Sparatore and D. Taramelli, *Bioorg. Med. Chem.*, 2005, **13**, 5338-5345.
3. T. J. Egan, R. Hunter, C. H. Kaschula, H. M. Marques, A. Misplon and J. Walden, *J. Med. Chem.*, 2000, **43**, 283-291.
4. T. J. Egan, *Mini Rev Med Chem*, 2001, **1**, 113-123.
5. C. H. Kaschula, T. J. Egan, R. Hunter, N. Basilico, S. Parapini, D. Taramelli, E. Pasini and D. Monti, *J. Med. Chem.*, 2002, **45**, 3531-3539.
6. L. M. B. Ursos and P. D. Roepe, *Med. Res. Rev.*, 2002, **22**, 465-491.
7. P. Olliaro, *Pharmacol. Ther.*, 2001, **89**, 207-219.
8. R. G. Ridley, W. Hofheinz, H. Matile, C. Jaquet, A. Dorn, R. Masciadri, S. Jolidon, W. F. Richter, A. Guenzi, M. A. Girometta, H. Urwyler, W. Huber, S. Thaithong and W. Peters, *Antimicrob. Agents Chemother.*, 1996, **40**, 1846-1854.
9. D. Y. De, F. M. Krogstad, F. B. Cogswell and D. J. Krogstad, *Am. J. Trop. Med. Hyg.*, 1996, **55**, 579-583.
10. D. Y. D. De, F. M. Krogstad, L. D. Byers and D. J. Krogstad, *J. Med. Chem.*, 1998, **41**, 4918-4926.
11. A. Ryckebusch, R. Deprez-Poulain, L. Maes, M. A. Debreu-Fontaine, E. Mouray, P. Grellier and C. Sergheraert, *J. Med. Chem.*, 2003, **46**, 542-557.
12. W. Hofheinz, C. Jaquet and S. Jolidon, *United States Patent*, 1997, **5596002**.
13. A. Golbraikh, M. Shen, Z. Y. Xiao, Y. D. Xiao, K. H. Lee and A. Tropsha, *Journal of Computer-Aided Molecular Design*, 2003, **17**, 241-253.
14. W. G. Hopkins, *A New View of Statistics - Log Transformation for Better Fits*, http://www.uq.edu.au/~hmrburge/stats/logtrans.html, Accessed 29/06/08.
15. R. Perkins, H. Fang, W. D. Tong and W. J. Welsh, *Environ. Toxicol. Chem.*, 2003, **22**, 1666-1679.
16. C. Bologa, T. Allu, M. Olah, M. Kappler and T. Oprea, *Journal of Computer-Aided Molecular Design*, 2005, **19**, 625-635.
17. R. Todeschini, V. Consonni, A. Mauri and M. Pavan, *DRAGON Web version*.
18. Fujitsu, *ADMEWORKS ModelBuilder*, http://www.fqs.pl/life_science/admeworks_modelbuilder.
19. R. Todeschini, V. Consonni and M. Pavan, *DRAGON 3.0*.
20. A. Berglund, M. C. D. Rosa and S. Wold, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 601-612.
21. M. Shen, A. LeTiran, Y. D. Xiao, A. Golbraikh, H. Kohn and A. Tropsha, *J. Med. Chem.*, 2002, **45**, 2811-2823.
22. OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, Paris, 2007.
23. D. J. Livingstone and E. Rahr, *Quant. Struct.-Act. Relat.*, 1989, **8**, 103-108.
24. R. Kohavi and G. H. John, *Artificial Intelligence*, 1997, **97**, 273-324.
25. L. Xu and W. J. Zhang, *Analytica Chimica Acta*, 2001, **446**, 477-483.
26. K. N. Berk, *SIAM Review*, 1992, **34**, 325-326.
27. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1989.
28. D. Rogers and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 854-866.
29. Y. Chun Wei, *PHAKISO - Pharmacokinetics In Silico*, http://www.phakiso.com/.
30. D. L. Massart, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier Science, 1997.
31. Y. Li, J. Liu, D. Pan and A. J. Hopfinger, *Toxicological Sciences*, 2005, **88**, 434-446.
32. H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry*, VCH, Weinheim, 1993.
33. P. Diaconis and B. Efron, *Sci.Am.*, 1983, **248**, 116-&.

34. R. D. Cramer, J. D. Bunce, D. E. Patterson and I. E. Frank, *Quant. Struct.-Act. Relat.*, 1988, **7**, 18-25.
35. A. Golbraikh and A. Tropsha, *J. Mol. Graph.*, 2002, **20**, 269-276.
36. R. Wehrens, H. Putter and L. M. C. Buydens, *Chemometrics Intell. Lab. Syst.*, 2000, **54**, 35-52.
37. A. Tropsha, P. Gramatica and V. K. Gombar, 10th International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences (QSAR 2002), Ottawa, Canada, 2002.
38. A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Comb. Sci.*, 2003, **22**, 69-77.
39. F. Lindgren, B. Hansen, W. Karcher, M. Sjostrom and L. Eriksson, 4th Scandinavian Symposium on Chemometrics, Lund, Sweden, 1995.
40. B. L. Podlogar, I. Muegge and L. J. Brice, *Current Opinion in Drug Discovery & Development*, 2001, **4**, 102-109.
41. A. Tropsha, *Molecular Informatics*, 2010, **29**, 476-488.
42. A. Golbraikh and A. Tropsha, *Molecular Diversity*, 2000, **5**, 231-243.
43. A. Golbraikh, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 414-425.
44. B. D. Hudson, R. M. Hyde, E. Rahr and J. Wood, *Quant. Struct.-Act. Relat.*, 1996, **15**, 285-289.
45. M. Snarey, N. K. Terrett, P. Willett and D. J. Wilton, *J. Mol. Graph.*, 1997, **15**, 372-385.
46. C. H. Reynolds, R. Druker and L. B. Pfahler, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 305-312.
47. L. Sachs, *Applied Statistics. A Handbook of Techniques.*, Springer-Verlag, 1984.
48. E. Benfenati, *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, Elsevier Science, 2007.
49. E. Sarıpınar, N. Geçen, K. Şahin and E. Yanmaz, *European Journal of Medicinal Chemistry*, 2010, **45**, 4157-4168.
50. I. B. Bersuker, *Curr. Pharm. Design*, 2003, **9**, 1575-1606.
51. Spartan, *Wavefunction, INC*, 2008.
52. *Spartan '08 - Tutorial and User's Guide*, Wavefunction, Inc., 2006-2009.
53. M. Wall, *GAlib: A C++ Library of Genetic Algorithm Components*, 1996.
54. *R 2.9.0*, http://www.r-project.org/.
55. D. L. Cooper, *Personal communication*, The University of Liverpool, 2009.
56. M. Hewitt, M. T. D. Cronin, J. C. Madden, P. H. Rowe, C. Johnson, A. Obi and S. J. Enoch, *Journal of Chemical Information and Modeling*, 2007, **47**, 1460.
57. J. R. Votano, M. Parham, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, Q. A. Xie and W. Tong, *Mutagenesis*, 2004, **19**, 365-377.
58. L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikstrom and S. Wold, *Multi- and Megavariate Data Analysis: Basic Principles and Applications*, Umetrics, 2006.
59. S. Wold, A. Ruhe, H. Wold and W. J. Dunn, *Siam Journal on Scientific and Statistical Computing*, 1984, **5**, 735-743.
60. D. L. Massart, *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier Science, 1997.
61. VCCLAB, *Virtual Computational Chemistry Laboratory*, http://www.vcclab.org.
62. R. Todeschini and P. Gramatica, *Quant. Struct.-Act. Relat.*, 1997, **16**, 120-125.
63. T. Hancock, R. Put, D. Coomans, Y. Vander Heyden and Y. Everingham, *Chemometrics Intell. Lab. Syst.*, 2005, **76**, 185-196.
64. S. Deshpande, V. R. Solomon, S. B. Katti and Y. S. Prabhakar, *Journal of Enzyme Inhibition and Medicinal Chemistry*, 2009, **24**, 94 - 104.
65. R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Wiley-VCH, 2000.
66. R. Put, Q. S. Xu, D. L. Massart and Y. Vander Heyden, *Journal of Chromatography A*, 2004, **1055**, 11-19.
67. J. Galvez, R. Garcia, M. T. Salabert and R. Soler, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 520-525.
68. J. Galvez, R. Garcia-Domenech, J. V. de Julian-Ortiz and R. Soler, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 272-284.

69. L. J. Soltzberg and C. L. Wilkins, *Journal of the American Chemical Society*, 1977, **99**, 439-443.
70. J. J. P. Stewart, *MoPAC*, Stewart Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net.
71. D. Young, *Computational Chemistry - A Practical Guide for Applying Techniques to Real World Problems*, 2001.
72. R. G. Pearson, *Journal of Chemical Sciences*, 2005, **117**, 369-377.
73. R. G. Pearson, *Journal of the American Chemical Society*, 1963, **85**, 3533-3539.
74. M. Schlitzer, *ChemMedChem*, 2007, **2**, 944-986.
75. W. F. Zheng and A. Tropsha, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 185-194.
76. H. J. H. MacFie, *Analytica Chimica Acta*, 1986, **186**, 345-345.
77. G. M. Maggiora and M. A. Johnson, *INTRODUCTION TO SIMILARITY IN CHEMISTRY*, John Wiley & Sons Inc, New York, 1990.
78. E. K. Wagner and M. J. Hewlett, *Basic Virology*, 2 edn., Blackwell Publishing, Malden, MA, 2004.
79. M. J. Tong, N. S. Elfarra, A. R. Reikes and R. L. Co, *N. Engl. J. Med.*, 1995, **332**, 1463-1466.
80. H. B. El-Serag and A. C. Mason, *N. Engl. J. Med.*, 1999, **340**, 745-750.
81. T. Poynard, V. Ratziu, Y. Benhamou, P. Opolon, P. Cacoub and P. Bedossa, *Best Pract. Res. Clin. Gastroenterol.*, 2000, **14**, 211-228.
82. J. Neyts, *Antiviral Res.*, 2006, **71**, 363-371.
83. T. Ohno, M. Mizokami, R. R. Wu, M. G. Saleh, K. Ohba, E. Orito, M. Mukaide, R. Williams and J. Y. N. Lau, *J. Clin. Microbiol.*, 1997, **35**, 201-207.
84. N. Sakamoto and M. Watanabe, *J. Gastroenterol.*, 2009, **44**, 643-649.
85. M. W. Fried, M. L. Shiffman, K. R. Reddy, C. Smith, G. Marinos, F. L. Goncales, D. Haussinger, M. Diago, G. Carosi, D. Dhumeaux, A. Craxi, A. Lin, J. Hoffman and J. Yu, *N. Engl. J. Med.*, 2002, **347**, 975-982.
86. M. P. Manns, J. G. McHutchison, S. C. Gordon, V. K. Rustgi, M. Shiffman, R. Reindollar, Z. D. Goodman, K. Koury, M. H. Ling, J. K. Albrecht and T. Int Hepatitis Interventional, *Lancet*, 2001, **358**, 958-965.
87. V. K. Rustgi, *Current Medical Research and Opinion*, 2009, **25**, 991-1002.
88. A. V. Stachulski, C. Pidathala, E. C. Row, R. Sharma, N. G. Berry, M. Iqbal, J. Bentley, S. A. Allman, G. Edwards, A. Helm, J. Hellier, B. E. Korba, J. E. Semple and J. F. Rossignol, *Journal of Medicinal Chemistry*, 2011, **54**, 4119-4132.
89. B. E. Korba, A. B. Montero, K. Farrar, K. Gaye, S. Mukerjee, M. S. Ayers and J. F. Rossignol, *Antiviral Res.*, 2008, **77**, 56-63.
90. J. F. Rossignol, *Expert Opinion on Drug Metabolism and Toxicology*, 2009, **5**, 667-674.
91. B. E. Korba, M. Elazar, P. Lui, J. F. Rossignol and J. S. Glenn, *Antimicrob. Agents Chemother.*, 2008, **52**, 4069-4071.
92. C. Yon, P. Viswanathan, J. F. Rossignol and B. Korba, *Antiviral Res.*, 2011, **91**, 233-240.
93. J. F. Rossignol, A. Elfert, Y. El-Gohary and E. B. Keeffe, *Gastroenterology*, 2009, **136**, 856-862.
94. J. F. Rossignol, A. Elfert and E. B. Keeffe, *J. Clin. Gastroenterol.*, 2010, **44**, 504-509.
95. B. Korba, M. Elazar, P. Liu, J. S. Glenn and J. F. Rossignol, *Hepatology*, 2008, **48**, 356A-356A.
96. A. V. Stachulski, C. Pidathala, E. C. Row, R. Sharma, N. G. Berry, A. S. Lawrenson, S. L. Moores, M. Iqbal, J. Bentley, S. A. Allman, G. Edwards, A. Helm, J. Hellier, B. E. Korba, J. E. Semple and J.-F. Rossignol, *Journal of Medicinal Chemistry*, 2011, **54**, 8670-8680.
97. R. Todeschini, V. Consonni and M. Pavan, *DRAGON 6.0*.
98. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361-1375.
99. R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137-&.
100. R. N. Jorissen and M. K. Gilson, *Journal of Chemical Information and Modeling*, 2005, **45**, 549-561.

101. H. Li, C. W. Yap, Y. Xue, Z. R. Li, C. Y. Ung, L. Y. Han and Y. Z. Chen, *Drug Development Research*, 2005, **66**, 245-259.
102. www.knime.org, *KNIME v2.3.3*, 2003-2011.

Chapter VII

Synthesis of Quinolone Antimalarials

7. Synthesis of Quinolone Antimalarials 385

7.1 Pyrroloquinolone Motif 386

7.2 Docking of the Pyrroloquinolone Motif 387

7.3 Introduction of Synthetic Techniques 389

7.3.1 Camps Cyclisation 390

7.3.2 Witkop Oxidation 390

7.3.3 Winterfeldt Oxidation 394

7.4 Synthesis of Novel Pyrroloquinolone Analogues 396

7.4.1 Alkylation Reaction 398

7.4.2 Wohl-Ziegler Bromination 400

7.4.3 Winterfeldt Oxidation 401

7.5 Biological Analysis 404

7.6 Topliss Scheme 406

7.7 Analysis of Results 409

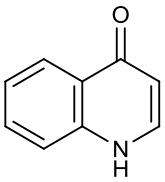
7.8 Docking of the Synthesised Pyrroloquinolones 410

7.9 Summary of Synthetic Study 413

7.10 References 414

7. Synthesis of Quinolone Antimalarials

Chemical synthesis is a vital part of the molecular design loop (fig. 4.1), and indeed all aspects of modern drug discovery. Medicinal chemists play a key role in optimising biological activity through SAR exploration and understanding, as well as contributing to the discovery of new chemotypes.[1] The importance of organic synthesis for the development of antimalarial treatments has already been discussed in Chapter I, with reference to lead candidates such as CQ.[2] As there are many potential pharmaceutical targets for the malaria parasite, synthetic work is vital in probing chemical space to find suitably active compounds. Several chemotypes have already been explored which target *Pf*bc$_1$, yet the hydroxynaphthoquinone compound ATOV[3-5] (fig. 1.17) is currently the only one in clinical use. Other promising candidates active against *Pf*bc$_1$ include pyridone compounds such as GW844520[6-8] (fig. 1.26), but due to the recent termination of development of this drug owing to unexpected cardiotoxicity, [9] it is clear that continued research efforts into new chemotypes is essential. One such chemotype which is currently undergoing much study is that of the basic quinolone core shown in figure 7.1.[10]

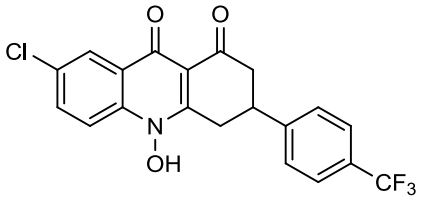


Quinolone

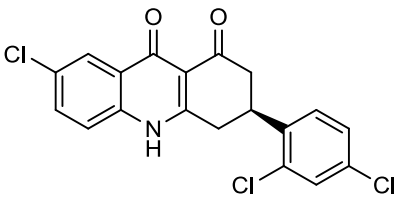
Fig. 7.1 Basic quinolone template.

Quinolone containing compounds have long been of interest as antimalarial agents against *Pf*bc$_1$,[11] with floxacrine and WR249685 (fig. 7.2) being just two examples.[9] These compounds both show haem binding as well as *Pf*bc$_1$ inhibitory properties. However, whilst floxacrine kills parasites via a haem mediated process similar to

that of CQ, WR249685 is a much weaker haem binder, and inhibits *Pf*bc$_1$ by binding selectively with the parasite Q_o site. In fact, WR249685 is around 5,000 fold more selective for the parasite bc$_1$ complex than that of human bc$_1$, and roughly 200 times more selective than ATOV. This is important as it reduces the potential for toxicity. This class of compound therefore has much potential for *Pf*bc$_1$ inhibition.



Floxacrine



WR249685 (S enantiomer)

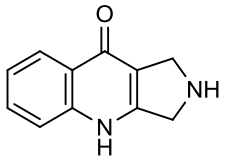
Fig. 7.2 Floxacrine and WR249685 quinolones.

The importance of quinolone compounds has already been discussed in Chapter V, in which several were docked into the *Pf*bc$_1$ active site in order to propose possible binding modes with relation to their biological activity.[12] The rationale behind the quinolone template lies in the understanding of its role in the Q_o active site. It appears to fix the position of the Rieske iron-sulphur protein, with the polar quinolone template forming key H-bond associations within the binding pocket,[10] and the long alkyl or aryl side chain residing in the hydrophobic pocket (i.e. fig. 5.22).

7.1 Pyrroloquinolone Motif

A SAR study around a particular chemotype or template is performed by making generally conservative alterations to a structure, based upon existing information or understandings. By combining the known antimalarial potential of the quinolone template (fig. 7.1) with new insight garnered from rigid tricyclic inhibitors such as floxacrine and WR249685, it was decided to expand upon the quinolone chemotype

and investigate, through periodic modifications, the pyrroloquinolone scaffold shown in figure 7.3.



Pyrroloquinolone

Fig. 7.3 Basic pyrroloquinolone scaffold.

Compounds containing the pyrroloquinolone scaffold, such as that in figure 7.4, have previously shown to be potent and selective PDE5 inhibitors. Compounds which inhibit PDE5 have found use as potential treatments for erectile dysfunction, and have been the subject of several SAR studies.[13, 14]

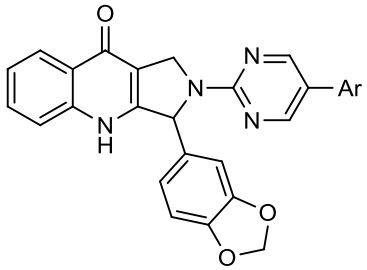


Fig. 7.4 Pyrroloquinolone lead structures known to inhibit PDE5. Ar denotes various aromatic side chains.

7.2 Docking of the Pyrroloquinolone Motif

Though the use of pyrroloquinolone containing compounds for the treatment of malaria has yet to be reported in the literature, the addition of a pyrrole ring to the quinolone template seemed like a natural progression of the quinolone SAR, particularly when considering the nature of the binding site. To offer additional support for the pyrroloquinolone template, a molecular docking study was undertaken. The Q_o active site is a compact and flat hole, opening into a large hydrophobic pocket. Rigid structures such as ATOV appear to fit nicely into this

active site,[15] so pyrroloquinolone compounds would be expected to replicate this mode of binding, and observe the H-bond interactions which are thought to be crucial for antimalarial activity.[16, 17]

The basic pyrroloquinolone template shown in figure 7.3 was first constructed and energy minimised using the *'Energy Minimisation Protocol'* as described in the Experimental Chapter, and then docked into the Q_o site using the *'Q_o Docking Protocol'*, also described in the Experimental Chapter. Constraints were applied such that poses were biased towards forming H-bonds with His181 and Glu272, and the crystallographic water molecule was allowed to translate and rotate within a radius of 2 Å. A total of 25 GA runs were performed, with the average GOLDScore and ChemScore values across the poses found to be 45.8 ± 1.8 and 28.9 ± 0.5 respectively. All the solutions were positioned in the centre of the Q_o pocket between residues Pro271 and Ile147. A strong H-bond was observed in all cases between the imidazole NH group of His181, and the carbonyl group of pyrroloquinolone. Most poses also showed potential to form the water mediated Glu272 interaction. Generally the solutions fell into one of two orientations, either the NH group of the pyrrole was directed towards the hydrophobic pocket, as in figure 7.5, or it was positioned facing the opposite direction. Either way, the template showed exciting potential as a bc_1 inhibitor, and by adding a side chain to the pyrrole nitrogen, it was hoped this would allow for additional hydrophobic or van der Waals contacts in the hydrophobic pocket of Q_o, ultimately strengthening the binding, and hopefully activity. This study represents how docking can be used to either support synthetic efforts, or drive them forward.

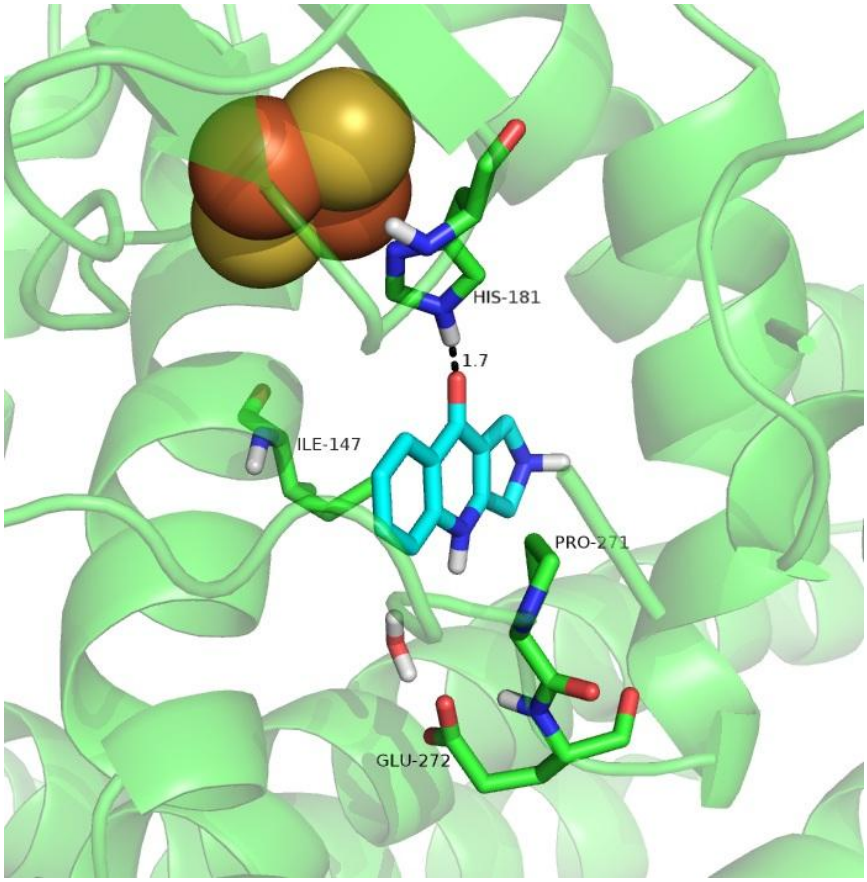


Fig. 7.5 Docking solution of pyrroloquinolone template (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The His181 H-bond with the pyrroloquinolone carbonyl group is clearly shown, as well as the potential for a water mediated interaction with Glu272. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

7.3 Introduction of Synthetic Techniques

The following sections discuss the synthesis and testing of a series of pyrroloquinolone compounds, with several side chains being substituted onto the nitrogen of the pyrrole ring, in order to allow for a small SAR analysis around the chemotype. A review of the literature for several of the synthetic procedures used will be followed by a detailed analysis of the results, together with interpretation of the data.

7.3.1 Camps Cyclisation

In 1899, a convenient route for the synthesis of quinolones was reported via Camps cyclisation, from simple acetyl-ortho-amidoacetophenone (fig. 7.6), where depending on the substituent's at the carbonyl groups, 2- or 4-quinolones were formed selectively.[18-22] Though reaction products can be depicted as quinolines, it is believed that the keto form (quinolone) predominates over its enol counterpart, both in solid state and in solution.[23]

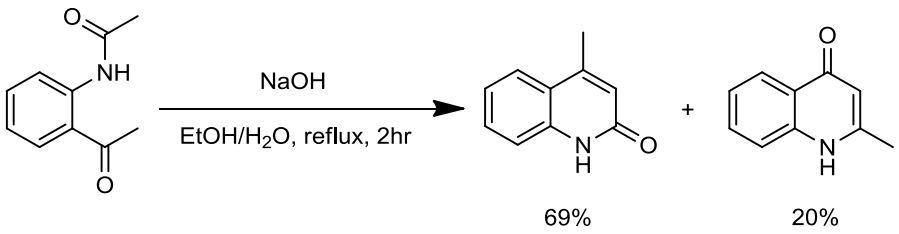


Fig. 7.6 Synthesis of 2- and 4-quinolone compounds via Camps cyclisation.

7.3.2 Witkop Oxidation

The Camps cyclisation received increased interest when it was found that the starting materials required for the reaction could be easily prepared through oxidative cleavage of the 2,3-double bond of substituted indoles, in a reaction termed Witkop oxidation.[24] The rational for the Witkop oxidation came from the understanding of an important intermediate in the biosynthetic oxidation of tryptophan in the liver, in which a dicarbonyl is formed from an indole (fig. 7.7).[25] Enzymes play a key role in this process in order to first oxidise the indole, and to then hydrolyse the amide to form kynurenine, an essential step in the biosynthesis of coenzyme NAD.[26, 27]

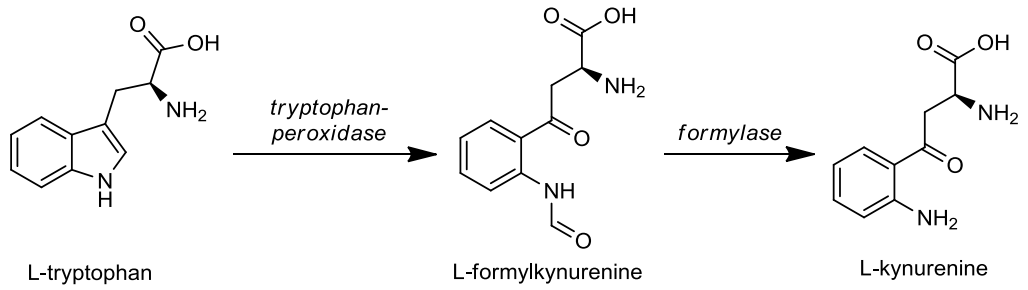


Fig. 7.7 Metabolism of tryptophan.

Through the study of ozonolysis for a series of indoles and indole containing natural products, it was possible to oxidise the indole double bond (fig. 7.8), and to even replicate the degradation of tryptophan to kynurenine.[28] The reaction was found to proceed via the formation of an ozonide, with subsequent rearrangement to the amide product.[29, 30]

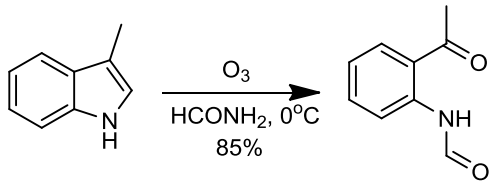


Fig. 7.8 Witkop oxidation with ozone.

Of particular importance was the combination of the ozonolysis reaction described by Witkop with a basic workup. This allowing for the conversion of several tetrahydroharman alkaloids into linear pyrroloquinolones, as depicted in figure 7.9.

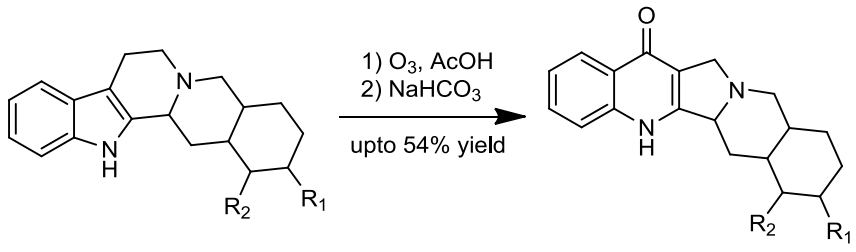


Fig. 7.9 Oxidation of tetrahydroharman alkaloids. R groups include hydrogen atoms, alcohol, carbonyl and ester functionality.

Alternative methods have been investigated for the oxidation of 2,3-disubstituted indole derivates, including the use of chromium trioxide (CrO_3). Though good

yields were reported for phenyl and methyl substituted indoles, the yields were poor for other alkyl substituent's (fig. 7.10).[31, 32]

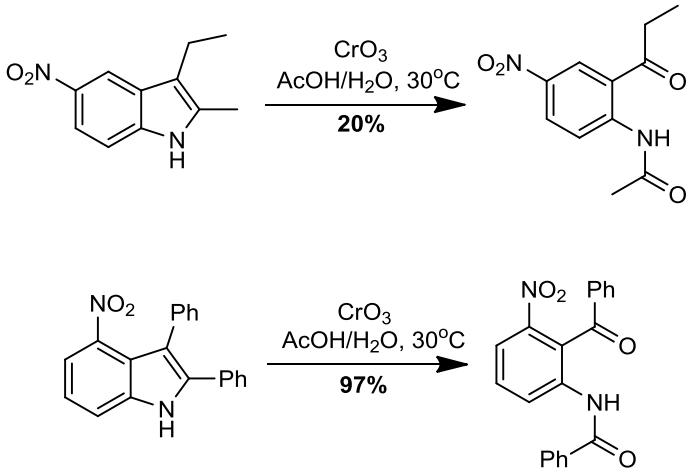


Fig. 7.10 Oxidation of indoles using chromium trioxide.

Witkop also investigated the use of oxygen in the presence of a platinum catalyst in order to oxidise the indole double bond, as illustrated by figure 7.11. The hydroperoxide intermediate rearranged to the dicarbonyl compound, with acidic conditions found to accelerate this rearrangement process.[33, 34] The mechanism of the hydroperoxide induced rearrangement was comparable to that of the ozonolysis reaction.

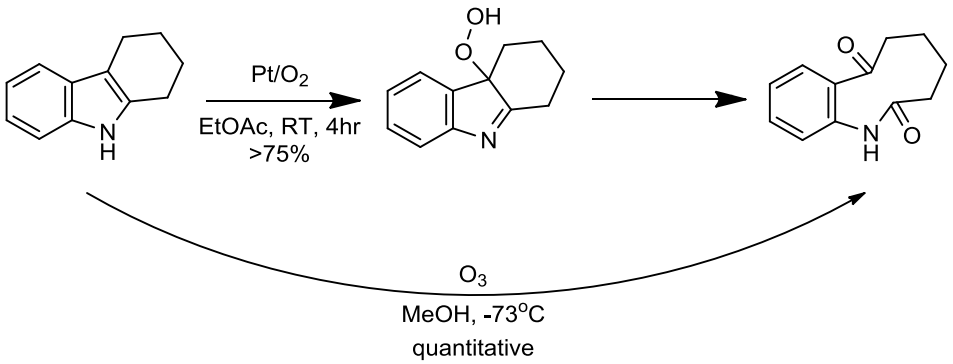


Fig. 7.11 Oxidation of indoles using Pt/O_2.

Through the study of a series of cycloalkylindoles it was found that the tendency to form the Camps cyclisation quinolone product was dependent on the ring size (fig.

7.12).[24, 35] Interestingly, the oxidation of the 8-membered ring led to the formation of a carbonyl compound instead of the dicarbonyl product.[36]

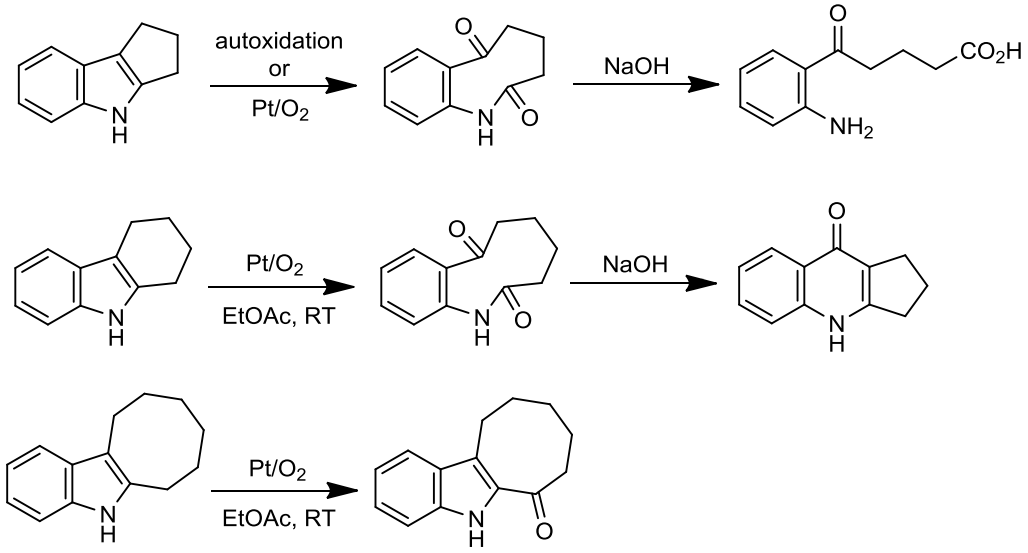


Fig. 7.12 Autoxidation/oxidation of cycloalkylindoles homologs.

During this study it was found that several of the investigated 2,3-disubstituted indoles formed their respective dicarbonyl product by autoxidation whilst standing on the shelf.[24, 37-39] A mechanism was subsequently proposed for oxidation which was consistent with that suggested for the Witkop oxidation (fig. 7.13).[40, 41] This mechanism shows the decomposition of the hydroperoxide bridge to form the dicarbonyl product.

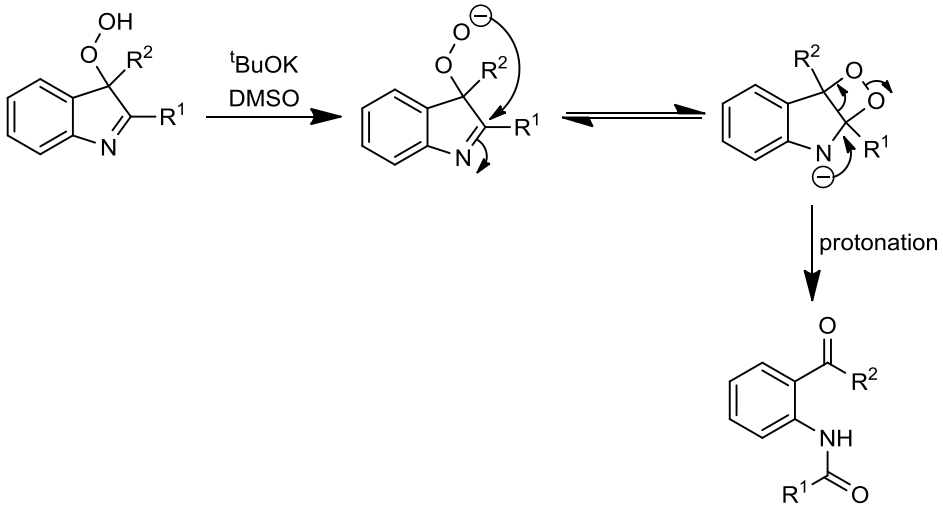


Fig. 7.13 Oxidation mechanism.

Additionally, sodium periodate ($NaIO_4$) has also provided a convenient oxidation reagent for the cleavage of the indole double bond, demonstrating excellent yields (fig 7.14).

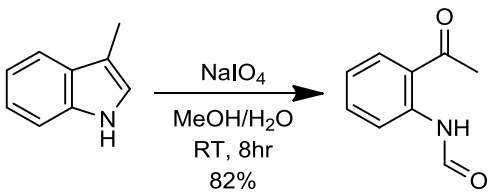


Fig. 7.14 Oxidation of indoles with $NaIO_4$.

7.3.3 Winterfeldt Oxidation

The sequential combination of Witkop oxidation and Camps cyclisation is commonly referred to as the Winterfeldt oxidation (fig. 7.15).[42] The Winterfeldt oxidation introduced NaH/O_2 and *t*-$BuOK/O_2$ as oxidation reagents, which owing to their basic nature, directly convert the dicarbonyl intermediates from the Witkop oxidation to their Camps cyclisation products.

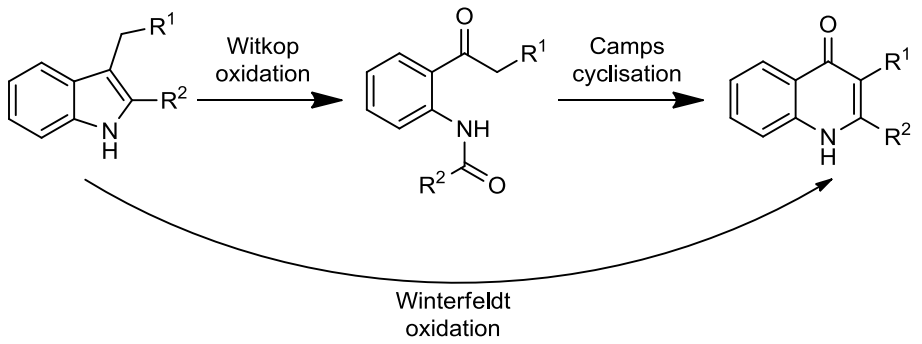


Fig. 7.15 Winterfeldt oxidation reaction.

Following suggestions that the natural product camptothecin, an anti leukaemia drug, is biosynthetically formed by oxidation of a corresponding carboline derivative and subsequent Camps cyclisation,[43] Winterfeldt exploited this strategy for the total synthesis of (±)-camptothecin (fig. 7.16).[44] What was of particular interest was that the selection of an inorganic base allowed the reaction to proceed directly from the

indole to the quinolone, without isolation of the dicarbonyl intermediate. In figure 7.16 the indole was oxidised using sodium hydride and oxygen in dimethylformamide (DMF), giving a yield of 75% for the pyrroloquinolone product.

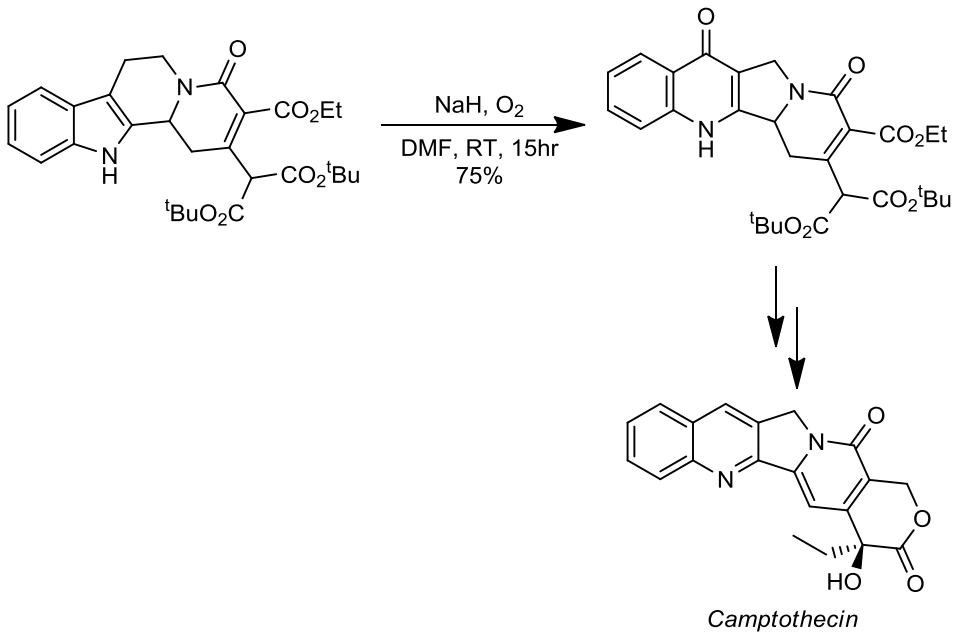


Fig. 7.16 Winterfeldt's synthesis of (±)-camptothecin.

Winterfeldt further investigated the direct conversion of a number of indoles to their quinolone counterpart.[42, 44, 45] One particular example is the oxidation of the natural product ajmalicine, which occurred with an excellent yield of 92% (fig. 7.17).

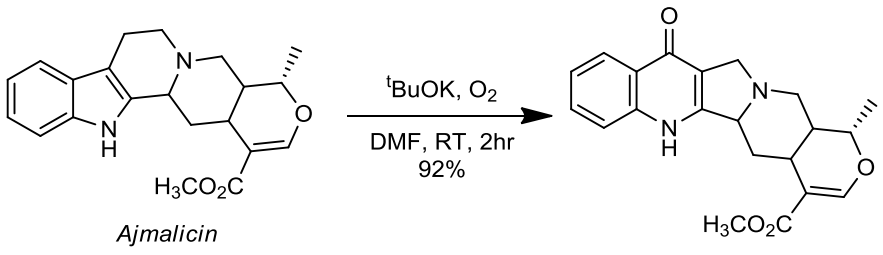


Fig. 7.17 Winterfeldt oxidation of indole alkaloids.

Additional studies using Winterfeldt's method have also been successfully applied for the preparation of pyrroloquinolones, as illustrated in figure 7.18.[46] Yields ranged between 40-92%, depending upon the side chains of the starting material.

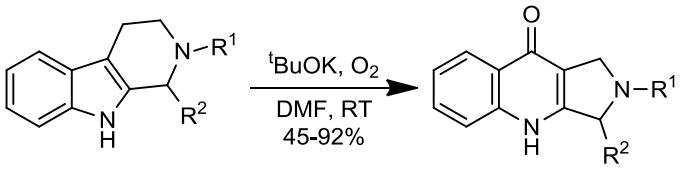


Fig 7.18 Synthesis of substituted pyrroloquinolones.

As was discussed earlier, pyrroloquinolone containing compounds have found use as PDE5 inhibitors for the treatment of male erectile dysfunction, many with activity in the picomolar (pM) range.[13, 14, 47] These series' of pyrroloquinolones were synthesised utilising the Winterfeldt oxidation procedure. It was also found that no racemisation occurred when stereoisomer's underwent Winterfeldt oxidation (fig. 7.19).[48]

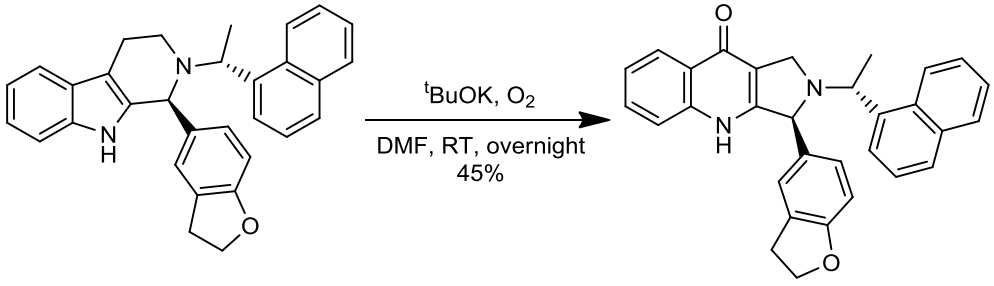


Fig 7.19 Synthesis of PDE5 inhibitors via Winterfeldt oxidation.

7.4 Synthesis of Novel Pyrroloquinolone Analogues

Given the previous successes of the Winterfeldt oxidation procedure it was used to synthesise the novel pyrroloquinolone analogues. In order to explore the SAR around this chemotype, a number of simple aromatic side chains were attached to the nitrogen of the pyrrole ring via a methyl linker (fig. 7.20).

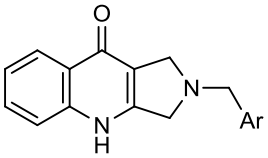


Fig. 7.20 Pyrroloquinolone template. Ar denotes aromatic substituents.

Figure 7.21 illustrates the two step procedure which was followed in order to synthesise the pyrroloquinolone compounds. The first step was to attach the required aromatic side chain to the tetrahydropyridoindole compound, which then underwent subsequent Winterfeldt oxidation, converting the indole to the pyrroloquinolone compound.

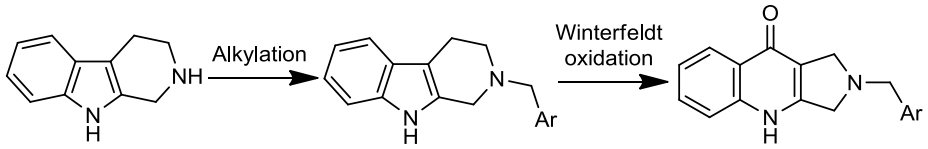


Fig. 7.21 Overall reaction scheme for the synthesis of novel pyrroloquinolone compounds.

The tetrahydropyridoindole compound was chosen as the starting material as it was commercially available. However, had this not been the case, it could have been synthesised via the Pictet Spengler reaction, which involves the condensation and ring closure of β-arylethylamines such as tryptamine, using an aldehyde or ketone in the presence of an acidic catalyst, usually whilst heating (fig. 7.22).[49, 50]

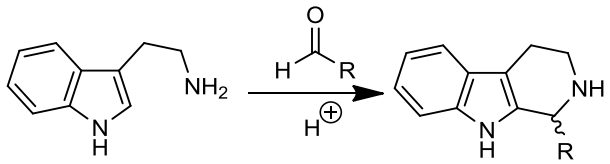


Fig. 7.22 Pictet Spengler reaction.

Figure 7.23 illustrates the mechanism of the Pictet Spengler reaction that begins with a Mannich reaction.[51] Under acidic conditions the amine attacks the aldehyde/ketone, after which the loss of water forms the iminium ion. It is the electrophilicity of this imine double bond which drives the electrophilic substitution at the 2-position,[52] after which the loss of a proton forms the final product.

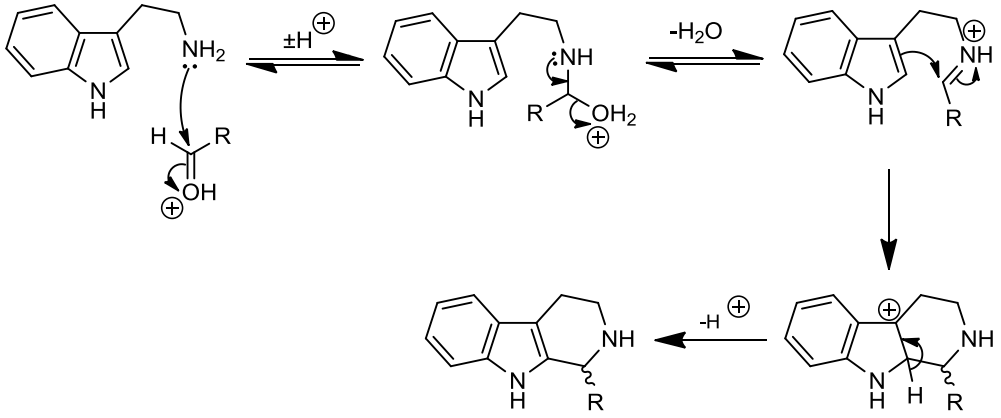


Fig. 7.23 Mechanism of the Pictet Spengler reaction.

7.4.1 Alkylation Reaction

Aromatic bromides were substituted onto the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (THPI) starting material via a fairly simple procedure, outlined by figure 7.24. One equivalent of THPI was dissolved in dry tetrahydrofuran (THF), together with two equivalents of triethylamine and 1.2 equivalents of the aromatic bromide in question. When the reaction was complete, the products were purified via flash column chromatography.

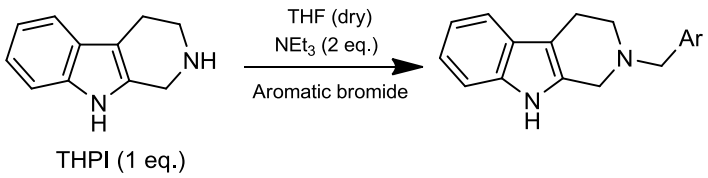
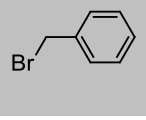
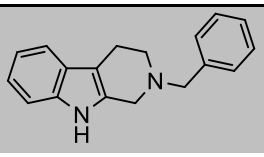
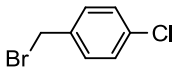
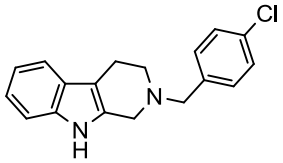
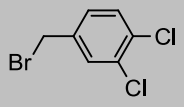
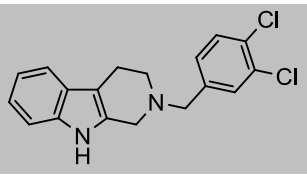
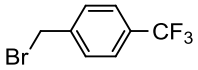
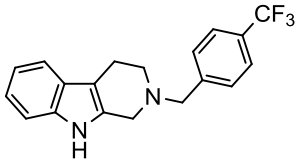
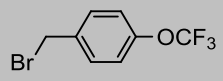
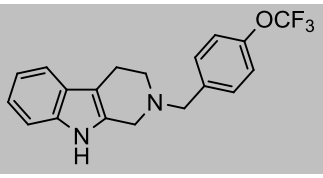
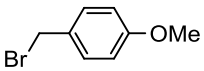
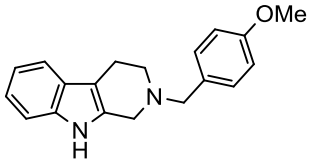
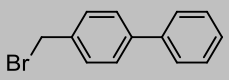
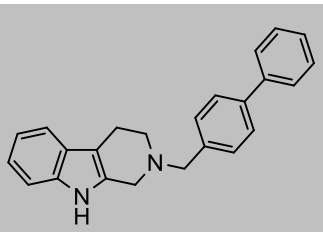


Fig. 7.24 General procedure for the alkylation reaction.

In total, seven aromatic bromides (**1-7**) underwent alkylation to THPI. The structures of these aromatics bromides, together with the products they formed and their respective yields are shown in table 7.1. The alkylation products have been labelled **8-14** and will be referred to as such from here on. The yields of these reactions were generally good and ranged from 46-86%. However, purification via flash column chromatography gave varying yields for the more polar compounds.

Table. 7.1 The substituted tetrahydropyridoindole products and their yields for the alkylation reactions.

Aromatic Bromide	ID	Tetrahydropyridoindole Product	ID	Appearance	Yield
	**1**		**8**	Pale yellow solid	66%
	**2**		**9**	Yellow solid	75%
	**3**		**10**	Pale orange solid	79%
	**4**		**11**	Pale yellow solid	86%
	**5**		**12**	Orange solid	86%
	**6**		**13**	Orange/yellow solid	53%
	**7**		**14**	Pale yellow solid	46%

The general mechanism for the alkylation reaction can be found in figure 7.25. An S_N2 type reaction occurs in which the lone pair of electrons on the amine nitrogen of THPI act as a nucleophile and attack the aromatic bromide at the electrophilic carbon. The proton of the resulting cation is then removed by triethylamine.

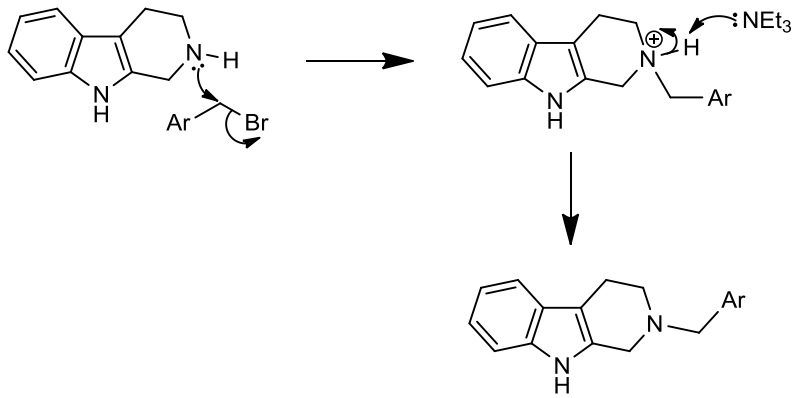


Fig. 7.25 Proposed mechanism for the alkylation reaction.

7.4.2 Wohl-Ziegler Bromination

Aromatic bromides **1-6** were all commercially available. However, to investigate how a simple extension of the side chain altered the *in vitro* behaviour of the compounds, aromatic bromide **7** had to be synthesised. This was done via the Wohl-Ziegler bromination of 4-phenyl toluene,[53] as it is a particularly useful reaction for the selective bromination of benzylic hydrogen's under neutral conditions.[54] The method which was used is outlined in figure 7.26, and involved dissolving 4-phenyltoluene in acetonitrile, together with 1.8 equivalents of *N*-bromosuccinimide (NBS), and 0.2 equivalents of 2,2'-azobisisobutyronitrile (AIBN) to act as a radical initiator. The reaction then proceeded under reflux until the 4-phenyltoluene had been consumed, with the product purified via flash column chromatography. The brominated product was isolated in an excellent yield of 89%.

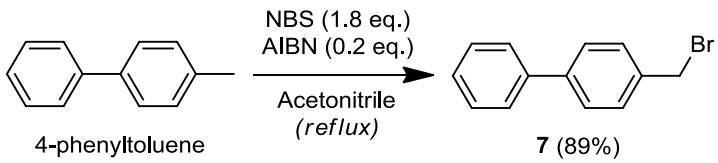


Fig. 7.26 Wohl-Ziegler bromination of 4-phenyltoluene.

The mechanism for this reaction, as depicted in figure 7.27, consists of initiation and propagation steps. Under heat and in the presence of the AIBN initiator, homolytic

cleavage of the NBS nitrogen-bromine bond occurs, allowing for the abstraction of a methyl hydrogen from 4-phenyltoluene to form a stabilised radical intermediate. This radical then reacts with NBS to form the bromo-substituted product **7**. **7** then underwent alkylation to THPI to form **14**.

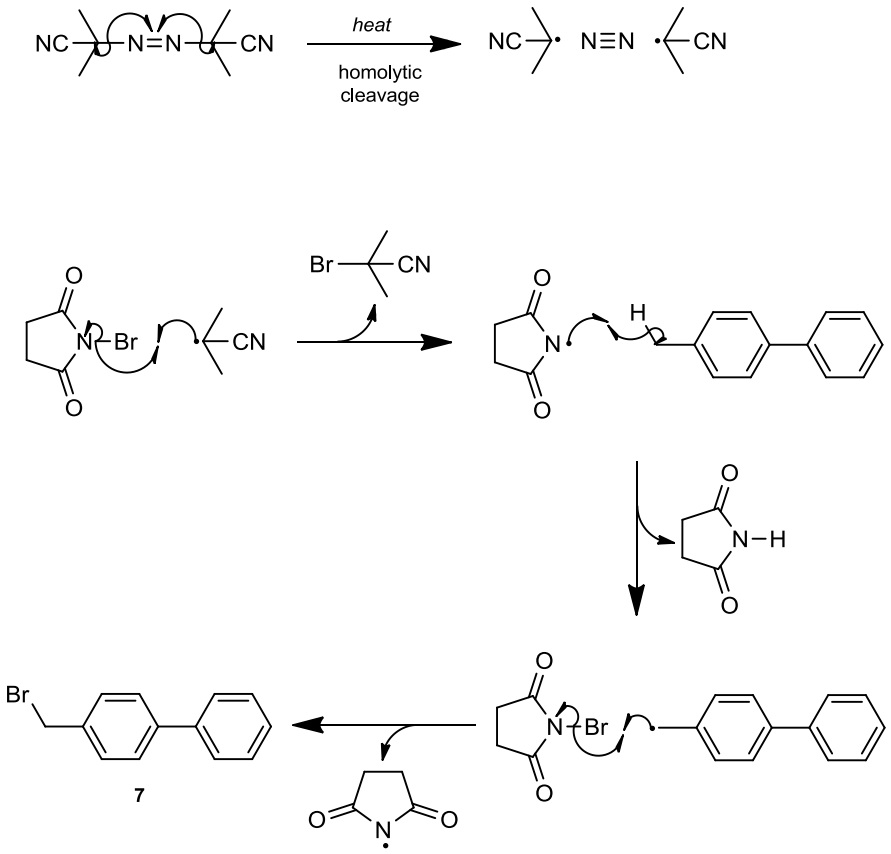


Fig. 7.27 Proposed mechanism for the Wohl-Ziegler bromination.

7.4.3 Winterfeldt Oxidation

The substituted tetrahydropyridoindoles were all converted to their pyrroloquinolone counterparts via the Winterfeldt oxidation procedure, outlined in figure 7.28. These conditions were shown to give good yields for a variety of substituent's.[20] This one pot synthesis allowed for the direct conversion of the tetrahydropyridoindoles **8-14** to their respective pyrroloquinolones. The general procedure involved dissolving a 1:1 mixture of the substituted tetrahydropyridoindole and potassium tertiary butoxide

in DMF. A constant stream of oxygen was bubbled through the solution until all of the tetrahydropyridoindole had been consumed. The mixture was then neutralised and an organic extraction and wash performed to isolate the product.

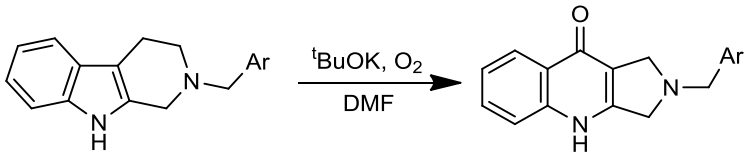
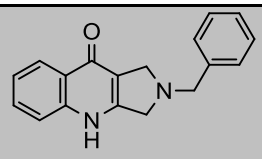
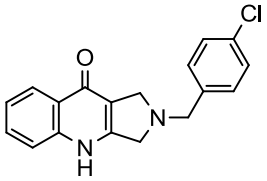
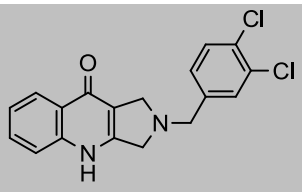
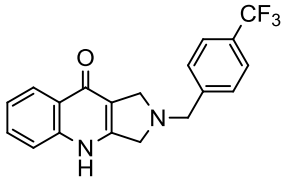
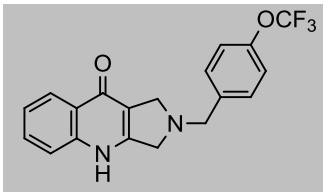
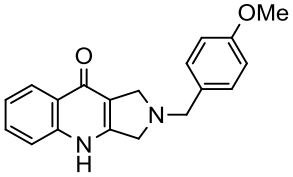
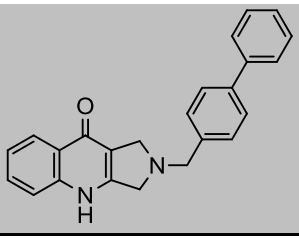


Fig. 7.28 General procedure for Winterfeldt oxidation reaction.

The tetrahydropyridoindoles **8-14** each underwent Winterfeldt oxidation to form the substituted pyrroloquinolone products **15-21**, shown in table 7.2. The yields of these reactions were variable and ranged between 24-79%. The poorer yields may be attributed to difficulties in extracting the product from the reaction mixture, owing to the limited solubility of the pyrroloquinolone compounds.

Table. 7.2 The substituted pyrroloquinolone products and their yields for the Winterfeldt oxidation reactions.

Pyrroloquinolone Product	ID	Appearance	Yield
	**15**	Yellow solid	33%
	**16**	Yellow solid	79%
	**17**	Yellow solid	48%
	**18**	Pale yellow solid	25%
	**19**	Yellow solid	32%
	**20**	Pale yellow solid	24%
	**21**	Yellow solid	41%

The proposed mechanism for the Winterfeldt oxidation can be found in figure 7.29, and begins with oxygen breaking the 2,3-indole double bond and forming a peroxide bridge.[42] Degradation of this peroxide bridge leads to the dicarbonyl/Witkop intermediate. The instability of this intermediate and the presence of base (^{t}BuOK)

drive forward the intramolecular aldol type reaction to form the stable six membered ring/Camps cyclisation product.[55]

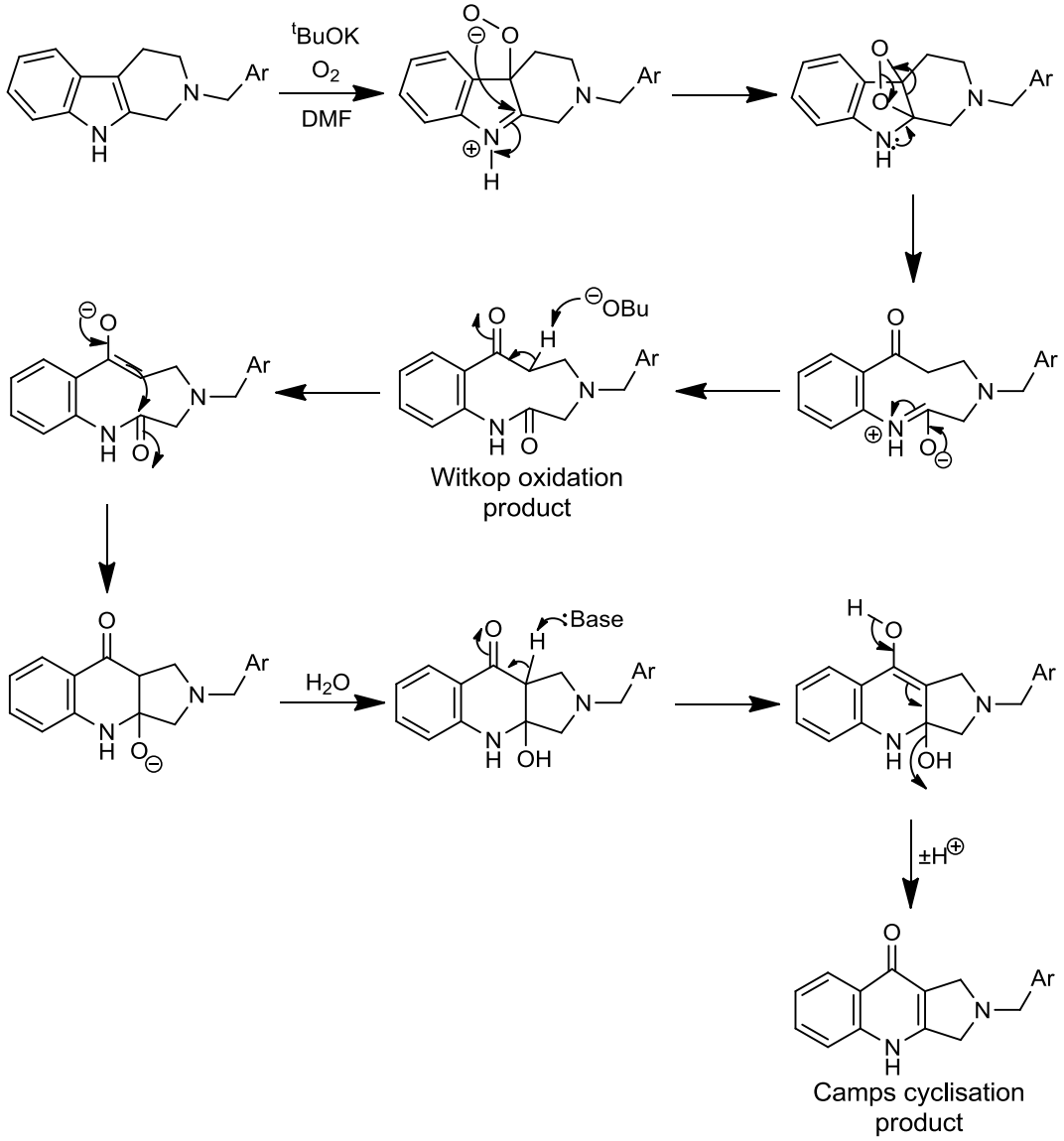


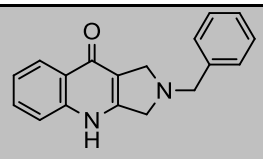
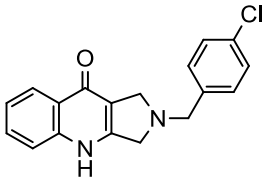
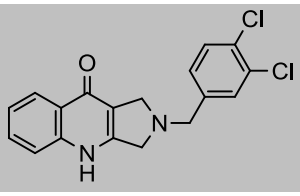
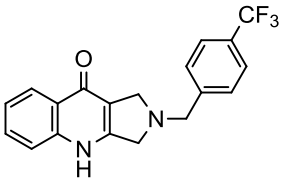
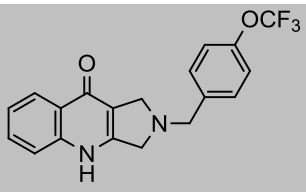
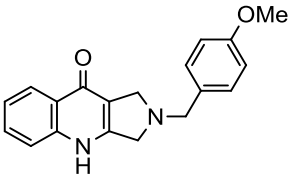
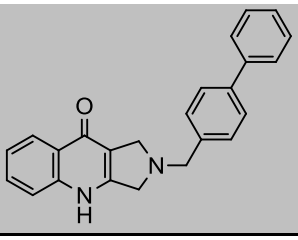
Fig. 7.29 Proposed mechanism for the Winterfeldt oxidation reaction.

7.5 Biological Analysis

±The substituted pyrroloquinolone compounds **15-21** were each tested *in vitro* using the *'Whole Cell Growth Inhibition Assay (3D7) Protocol'* as described in the Experimental Chapter. This assay was chosen as it would provide a quantitative measure of the activity of each of the analogues against malaria, allowing for

comment to be drawn as to the potential SAR of the series. The 3D7 whole cell IC_{50} values are reported in table 7.3. At the time of reporting the testing had only been performed once.

Table. 7.3 3D7 whole cell inhibition IC_{50} values for compounds **15-21**.

Pyrroloquinolone Product	ID	3D7 Whole Cell IC_{50} (nM)
	**15**	136
	**16**	552
	**17**	75
	**18**	352
	**19**	179
	**20**	1020
	**21**	774

As can be seen the biological testing results span a wide range of activities, from as few as 75 nM to 1.02 μM. This allowed for some interesting observations to be made.

7.6 Topliss Scheme

The Topliss scheme provides a means for navigating through the SAR of a chemotype, in order to maximise the chances of synthesising the most potent compounds in a series as early as possible.[1, 56] The procedure is particularly useful when compounds are fairly simple to synthesise, yet there is a considerable time lag in acquiring activity data. A Topliss scheme takes the form of a decision tree flow diagram and provides a route for selecting the most promising substituent to make. The Topliss schemes have been designed by considering how the hydrophobicity and electronic factors of various substituent's affect a compounds behaviour, and are designed such that optimum substituent's can be found as efficiently as possible.

The hydrophobicity of a drug is crucial to how easily it crosses the cell membrane, and may also be important in receptor interactions.[1] Changing the substituent's on a drug may well have significant effects on its hydrophobic character, and therefore its biological activity. The partition coefficient of a compound can be calculated theoretically by knowing the contribution that the various substituent's make. This overall contribution is known as the substituent hydrophobicity constant (π), and is a measure of how hydrophobic a substituent is relative to hydrogen. If π has a positive value then the substituent is more hydrophobic than hydrogen, if it is negative then it is less hydrophobic. The π values are characteristic for the substituent and can be used to calculate how the partition coefficient (log *P*) of a drug would be affected if these substituent's were present.

The electronic effects of various substituent's will have an impact on a drugs ionisation or polarity, which may in turn have an effect on how easily a drug can pass through cell membranes, or how strongly it can interact with a binding site.[1] It is therefore useful to measure the electronic effect of a substituent. The Hammett substitution constant (σ) provides a measure of the electron withdrawing or donating ability of a particular substituent. A positive value indicates an electron withdrawing group such as nitro, which has a stabilising influence on the ionised form of a compound. A negative Hammett constant value however indicates an electron donating group such as an alkyl substituent, which shifts the equilibrium the other way (fig. 7.30).

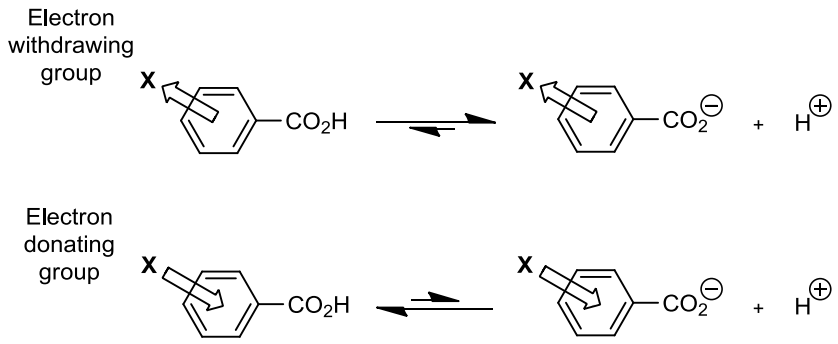


Fig. 7.30 Position of equilibrium depending on substituent group X.

The Topliss scheme in figure 7.31 can be used when substitutions are being made to an aromatic side chain. It assumes that the lead compound has a unsubstituted aromatic ring, whose biological activity is already known. The next analogue in the scheme is the 4-chloro derivative. The chloro substituent is more hydrophobic and electron withdrawing than hydrogen, and so has positive π and σ values. When the analogue has been synthesised it can be tested, after which there are three possibilities as to which branch of the Topliss scheme to follow. The analogue will be one of the following in comparison to the unsubstituted parent compound: less active (L); equally active (E); more active (M). Generally, if the activity increases

for the analogue then the M branch is followed, if it remains the same the E branch is followed, and if it decreases the L branch is followed. Then the next analogue along this branch is synthesised and the process repeated.

Fig. 7.31 Topliss scheme for aromatic substituent's (J. G. Topliss, *Journal of Medicinal Chemistry*, 1972, **15**, 1006-1011.)

If the activity increases with the 4-chloro derivative, then as the chloro group has positive π and σ values, this implies that these properties are important for biological activity. Thus, if π and σ are both important, an additional chloro group should increase the biological activity further. If it does, then the substituent's can be varied further to increase these values. If however it doesn't, then an unfavourable steric interaction or excessive hydrophobicity is indicated.

If the opposite to the above is true for the 4-chloro derivative, and the activity decreases, this suggests that negative π and σ values are important for biological activity, or that *para* substituent's are sterically unfavourable. It is first assumed that an unfavourable σ effect is the most likely reason for the reduced activity, and so the next substituent is one with a negative σ effect (i.e. OMe). If activity improves then further changes are suggested to test the relative importance of the σ and π factors.

However, if OMe fails to improve activity, then it is assumed unfavourable steric factors are at work and the next substitution occurs in the *meta* position.

7.7 Analysis of Results

The biological results of compounds **15-21** will now be interpreted with reference to the Topliss scheme (fig. 7.31). The parent compounds **15**, that is the pyrroloquinolone structure with an unsubstituted benzene ring attached to the pyrrole nitrogen via a methyl linker, reported an activity of 136 nM. Despite being the first analogue in the series, this proved to be a very promising start, demonstrating good nM activity. The initial modification (based on Topliss) was to substitute the hydrogen in the *para* position of the benzene ring with a chloro group (**16**). This was actually shown to decrease the biological activity to 552 nM, so in line with the Topliss scheme the next substitution to be made was to consider an OMe group (**20**). However, this greatly reduced the activity further, to only 1.02μM (least active analogue in the series). This may suggest that *para* substituent's are sterically unfavourable, and that substituent's with a negative σ value (electron donating groups) have a detrimental effect on activity. However, before investigating this branch further, it was decided to first investigating the effect of substituent's with increased π and σ values. Despite the 4-chloro analogue decreasing activity in comparison to the unsubstituted compound, the inclusion of an addition chloro group in the *meta* position (**17**) gave the most potent derivative, with an excellent activity of 75 nM. This did indeed suggest that electron withdrawing groups were preferable for this chemotype, as were more hydrophobic substitution patterns. The availability of reagents did limit the exploration of the Topliss scheme slightly, however, additional derivatives were synthesised and led to some interesting observations. As

predicted by Topliss, the substitution of a CF_3 group in the *para* position (**18**) did decrease the activity compared to the dichloro (**17**) derivative, with an activity of 352 nM. Yet the inclusion of a methoxy linker (OCF_3, **19**) led to a compound with an activity of 179 nM, similar to that of the parent molecule (**15**), but interestingly this had a more favourable solubility profile. An extension of the Topliss scheme was to lengthen the side chain by substituting into the *para* position an additional phenyl group (**21**). This was however shown to decrease the activity (774 nM). This again may be to due to unfavourable steric interactions at this position.

The dichloro compound (**17**) was the most potent compound in the series, and thus represents the lead candidate. Future testing and validation of its activity will allow for a more in depth analysis of its potential as an antimalarial drug. There is still however much scope for consideration with regard to the pyrroloquinolone template. The Topliss scheme suggests that 3-CF_3-4-Cl and 3-CF_3-4-NO_2 substitutions may lead to more active derivatives, but it may also be prudent to investigate modifications and substitutions in the *meta* position, as would similar exploration around the SAR of the biphenyl derivative (**21**). One potential limitation of these pyrroloquinolone compounds may be their poor solubility in most solvents, so chemical modification to improve this may be a rewarding avenue of investigation.

7.8 Docking of the Synthesised Pyrroloquinolones

To further validate the potential of the synthesised pyrroloquinolone compounds as *Pf*bc_1 inhibitors, structures **15-21** were each docked into the Q_o site of the yeast bc_1 complex (PDB accession code 3CX5)[57] using the *'Q_o Docking Protocol'*, following the energy minimisation of their structures using the *'Energy Minimisation Protocol'*. The average GOLDScore and ChemScore values across the 10 GA runs

are shown in table 7.4. It was hoped that from these results, comment could be drawn with regard to their binding orientation in the active site, as well as any potential trends between docking scores and *in vitro* activity.

Table. 7.4 Docking results for the synthesised pyrroloquinolone compounds at the Qo site.

	15	**16**	**17**	**18**	**19**	**20**	**21**
Average GOLDScore	58.5 ± 1.6	62.3 ± 0.5	61.2 ± 0.8	58.9 ± 0.5	62.1 ± 1.2	61.6 ± 1.1	56.1 ± 3.0
Average ChemScore	38.0 ± 2.2	39.5 ± 0.2	40.4 ± 2.3	36.9 ± 1.6	36.0 ± 0.9	36.8 ± 2.5	41.9 ± 1.0

As can be seen the docking scores were all fairly similar across the seven compounds, and given the ranging activity values for these compounds (table 7.3), there appears to be no linear relationship between fitness scoring and activity. This is not surprising however, given that they all share the central pyrroloquinolone chemotype and vary only in their side chains. Furthermore, the similar fitness scores across the series can be further explained through consideration of their docking poses. The compounds all docked similar to one another, with only slight movements in the location of the pyrroloquinolone core in the Q_o pocket, and their respective side chains in the hydrophobic pocket. These binding poses were very similar to that of the native binding ligand SMA (see fig. 5.5).

The docking pose of **17**, the most active compound in the pyrroloquinolone series, is shown in figure 7.32. This docking pose is representative of the entire series, with the carbonyl oxygen of the pyrroloquinolone core forming a strong H-bond with His181, as well as a water mediated H-bonding network between Glu272, and the amine group. The presence of these interactions offers a strong argument to account for their antimalarial activity, rationalising the observed activity for this series of compounds.

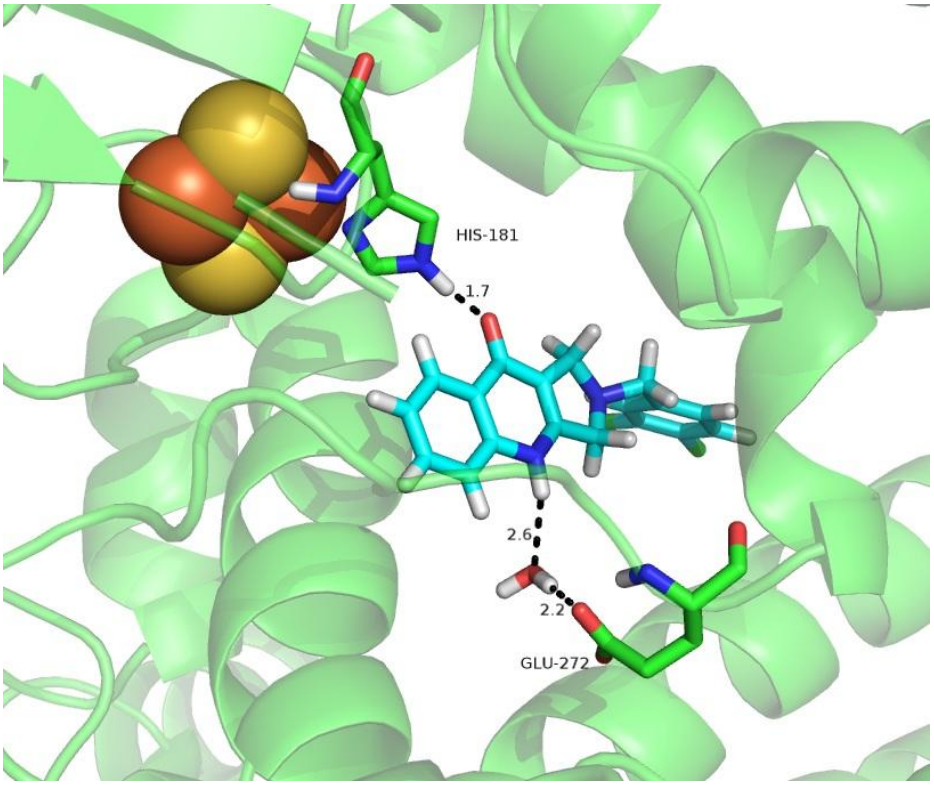


Fig. 7.32 Docking pose of **17** (shown in blue) in the Q_o pocket of the yeast cytochrome bc_1 complex (3CX5). The two key interactions, His181 and the water mediate Glu272 H-bond network are clearly illustrated. The yeast cytochrome b polypeptide backbone is represented in green, with the [2Fe2S] cluster of the Rieske protein represented as spheres (sulphur: gold, iron: orange). H-bonds are indicated by black lines.

The inclusion of the aromatic side chain also appears to have strengthened the docking when compared to the previous study of just the pyrroloquinolone core alone (fig. 7.5; respective GOLDScore and ChemScore values of 45.8 ± 1.8 and 28.9 ± 0.5). This may be due to an increased number of van der Waals and/or hydrophobic interaction in the hydrophobic pocket, or perhaps additional stability of the compounds in the Q_o pocket.

Clearly the docking protocol is unable to distinguish quantitatively between good and poorly active compounds. However, it can nicely predict possible binding orientations for known active compounds, each of which observe interactions known to be crucial for activity.[58]

7.9 Summary of Synthetic Study

Synthetic work has led to the rational design of a novel chemotype active against the malaria parasite. A SAR investigation of the pyrroloquinolone template has shown it to have much potential and scope, with molecular docking offering further support for possible Q_o binding and therefore *Pf*bc_1 inhibition. Subsequent modifications of the pyrroloquinolone side chain around the Topliss scheme may yield more potent hits, but it will be crucial to remain mindful of solubility issues with regard to these compounds, and work should therefore be performed to improve upon this.

7.10 References

1. G. L. Patrick, *An Introduction to Medicinal Chemistry*, Oxford University Press, 2005.
2. M. Schlitzer, *ChemMedChem*, 2007, **2**, 944-986.
3. M. Fry and M. Pudney, *Biochem. Pharmacol.*, 1992, **43**, 1545-1553.
4. M. W. Mather, E. Darrouzet, M. Valkova-Valchanova, J. W. Cooley, M. T. McIntosh, F. Daldal and A. B. Vaidya, *J. Biol. Chem.*, 2005, **280**, 27458-27465.
5. J. Krungkrai, S. R. Krungkrai, N. Suraveratum and P. Prapunwattana, *Biochem. Mol. Biol. Int.*, 1997, **42**, 1007-1014.
6. H. Xiang, J. McSurdy-Freed, G. S. Moorthy, E. Hugger, R. Bambal, C. Han, S. Ferrer, D. Gargallo and C. B. Davis, *J. Pharm. Sci.*, 2006, **95**, 2657-2672.
7. C. L. Yeates, J. F. Batchelor, E. C. Capon, N. J. Cheesman, M. Fry, A. T. Hudson, M. Pudney, H. Trimming, J. Woolven, J. M. Bueno, J. Chicharro, E. Fernandez, J. M. Fiandor, D. Gargallo-Viola, F. G. de las Heras, E. Herreros and M. L. Leon, *Journal of Medicinal Chemistry*, 2008, **51**, 2845-2852.
8. A. R. Crofts, B. Barquera, R. B. Gennis, R. Kuras, M. Guergova-Kuras and E. A. Berry, *Biochemistry*, 1999, **38**, 15807-15826.
9. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
10. R. W. Winter, J. X. Kelly, M. J. Smilkstein, R. Dodean, D. Hinrichs and M. K. Riscoe, *Exp. Parasitol.*, 2008, **118**, 487-497.
11. R. W. Winter, J. X. Kelly, M. J. Smilkstein, R. Dodean, G. C. Bagby, R. K. Rathbun, J. I. Levin, D. Hinrichs and M. K. Riscoe, *Exp. Parasitol.*, 2006, **114**, 47-56.
12. R. Cowley, S. Leung, N. Fisher, M. Al-Helal, N. G. Berry, A. S. Lawrenson, R. Sharma, A. E. Shone, S. A. Ward, G. A. Biagini and P. M. Oneill, *MedChemComm*, 2012.
13. Z. H. Sui, J. H. Guan, M. J. Macielag, W. Q. Jiang, S. Y. Zhang, Y. H. Qiu, P. Kraft, S. Bhattacharjee, T. M. John, D. Haynes-Johnson and J. Clancy, *Journal of Medicinal Chemistry*, 2002, **45**, 4094-4096.
14. J. C. Lanter, Z. H. Sui, M. J. Macielag, J. J. Fiordeliso, W. Q. Jiang, Y. H. Qiu, S. Bhattacharjee, P. Kraft, T. M. John, D. Haynes-Johnson, E. Craig and J. Clancy, *Journal of Medicinal Chemistry*, 2004, **47**, 656-662.
15. V. Barton, N. Fisher, G. A. Biagini, S. A. Ward and P. M. O'Neill, *Curr. Opin. Chem. Biol.*, 2010, **14**, 440-446.
16. B. L. Trumpower, *Biochim. Biophys. Acta-Bioenerg.*, 2002, **1555**, 166-173.
17. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
18. R. Camps, *Arch. Pharm.*, 1899, **237**, 659-691.
19. R. Camps, *Ber. Deutsch. Chem. Ges.*, 1899, 3228-3234.
20. M. Mentel and R. Breinbauer, *Curr. Org. Chem.*, 2007, **11**, 159-176.
21. M. A. Prezent and V. A. Dorokhov, *Russ. Chem. Bull.*, 2003, **52**, 2454-2456.
22. R. H. Manske, *Chemical Reviews*, 1942, **30**, 113-144.
23. C. P. Jones, K. W. Anderson and S. L. Buchwald, *The Journal of Organic Chemistry*, 2007, **72**, 7968-7973.
24. B. Witkop, J. B. Patrick and M. Rosenblum, *Journal of the American Chemical Society*, 1951, **73**, 2641-2647.
25. A. Ek, H. Kissman, J. B. Patrick and B. Witkop, *Experientia*, 1952, **8**, 36-40.
26. W. E. Knox and A. H. Mehler, *J. Biol. Chem.*, 1950, **187**, 419-430.
27. A. H. Mehler and W. E. Knox, *J. Biol. Chem.*, 1950, **187**, 431-438.
28. B. Witkop, *Justus Liebigs Ann. Chem.*, 1944, **556**, 103-114.
29. B. Witkop and J. B. Patrick, *Journal of the American Chemical Society*, 1952, **74**, 3855-3860.
30. B. Witkop and J. B. Patrick, *Journal of the American Chemical Society*, 1952, **74**, 3861-3866.
31. K. Schofield and R. S. Theobald, *Journal of the Chemical Society (Resumed)*, 1949, 796-799.
32. K. Schofield and R. S. Theobald, *Journal of the Chemical Society (Resumed)*, 1950, 1505-1509.
33. B. Witkop and J. B. Patrick, *Journal of the American Chemical Society*, 1951, **73**, 2188-2195.

34. B. Witkop and J. B. Patrick, *Journal of the American Chemical Society*, 1951, **73**, 2196-2200.
35. B. Witkop and S. Goodwin, *Journal of the American Chemical Society*, 1953, **75**, 3371-3376.
36. E. Leete, *Journal of the American Chemical Society*, 1961, **83**, 3645-3647.
37. R. J. S. Beer, T. Donavanik and A. Robertson, *Journal of the Chemical Society (Resumed)*, 1954, 4139-4142.
38. R. J. S. Beer, L. McGrath and A. Robertson, *Journal of the Chemical Society (Resumed)*, 1950, 2118-2126.
39. R. J. S. Beer, L. McGrath and A. Robertson, *Journal of the Chemical Society (Resumed)*, 1950, 3283-3286.
40. F. McCapra and Y. C. Chang, *Chemical Communications (London)*, 1966, 522-523.
41. F. McCapra, D. G. Richardson and Y. C. Chang, *Photochemistry and Photobiology*, 1965, **4**, 1111-1121.
42. E. Winterfeldt, *Annalen Der Chemie-Justus Liebig*, 1971, **745**, 23-30.
43. E. Wenkert, K. G. Dave, R. G. Lewis and P. W. Sprague, *Journal of the American Chemical Society*, 1967, **89**, 6741-6745.
44. M. Boch, Winterfe.E, J. M. Nelke, H. Radunz, D. Pike and T. Korth, *Chem. Ber.-Recl.*, 1972, **105**, 2126-&.
45. J. Warneke and Winterfe.E, *Chem. Ber.-Recl.*, 1972, **105**, 2120-&.
46. J. F. Carniaux, C. KanFan, J. Royer and H. P. Husson, *Tetrahedron Lett.*, 1997, **38**, 2997-3000.
47. W. Q. Jiang, Z. H. Sui, M. J. Macielag, S. P. Walsh, J. J. Fiordeliso, J. C. Lanter, J. H. Guan, Y. H. Qiu, P. Kraft, S. Bhattacharjee, E. Craig, D. Haynes-Johnson, T. M. John and J. Clancy, *Journal of Medicinal Chemistry*, 2003, **46**, 441-444.
48. W. Q. Jiang, Z. H. Sui and X. Chen, *Tetrahedron Lett.*, 2002, **43**, 8941-8945.
49. A. Pictet and T. Spengler, *Berichte der deutschen chemischen Gesellschaft*, 1911, **44**, 2030-2036.
50. W. M. Whaley and T. R. Govindachari, *Organic Reactions*, 1951, **6**, 151-190.
51. J. Clayden, N. Greeves, S. Warren and P. Wothers, *Organic Chemistry*, Oxford University Press, 2001.
52. E. D. Cox and J. M. Cook, *Chemical Reviews*, 1995, **95**, 1797-1842.
53. C. Djerassi, *Chemical Reviews*, 1948, **43**, 271-317.
54. H. Togo and T. Hirai, *Synlett*, 2003, 702-704.
55. M. Mentel, M. Peters, J. Albering and R. Breinbauer, *Tetrahedron*, 2011, **67**, 965-970.
56. J. G. Topliss, *Journal of Medicinal Chemistry*, 1972, **15**, 1006-1011.
57. S. R. N. Solmaz and C. Hunte, *J. Biol. Chem.*, 2008, **283**, 17542-17549.
58. L. Esser, B. Quinn, Y. F. Li, M. Q. Zhang, M. Elberry, L. Yu, C. A. Yu and D. Xia, *Journal of Molecular Biology*, 2004, **341**, 281-302.

Conclusions & Future Work

Malaria is a devastating disease, and continued efforts are required to develop new therapeutic options to both treat and prevent the disease. The work described within this thesis explores possible solutions for the discovery of novel antimalarial compounds using a host of methods spanning multiple specialities. Drug discovery in itself is multidisciplinary, requiring collaboration and expertise in many areas. Chemoinformatics and chemical synthesis have both been used here to discover new chemotypes active against malaria, with these results validated through biological screening.

Following an extensive introduction of malaria and modern drug discovery methods in Chapter I, Chapter II discussed a host of LBVS methods and their application. Using the chemical structures of 19 compounds with known *Pf*bc$_1$ activities, six methods were applied to the ZINC lead like library of compounds, generating a diverse range of hits. In Chapter III these hits were merged, and each molecule scored based on its frequency across the methods, and also their physicochemical properties. Consensus analysis was used to place emphasis onto compounds identified across several methods, as these had received the most support from virtual screening. Compounds were filtered to remove those with unfavourable chemical properties, or whose structure contained known toxicophores as reported across the literature. A final selection of 19 compounds was made, and these purchased and tested as described in Chapter IV, using several different bioassays. 5 of the compounds reported single digit μM IC_{50} values, with each containing novel structural chemotypes, such as an isoalloxazine ring, or a pyrrolopyrimidine-2,4-dione motif. The lead candidate contained a benzothiazole core, and reported an IC_{50} value of 4.53 ± 1.86 μM. Additional testing showed the compounds to have

little or no inhibition of bovine bc_1, which is highly promising as bovine bc_1 inhibition has been shown to be indicative of cardiotoxicity in humans.

Future work with regard to these 5 active hits will include validation of their site of action, as these compounds were selected based on their potential as *Pf*bc_1 inhibitors. This will require testing against the *Pf*bc_1 bioassay. This will be performed as and when the parasite can be cultured in the sufficient amount required for testing. Further to this, the most promising candidates will form the basis of the next iteration of the molecular design loop. In particular, the benzothiazole chemotype will become the focus of a SAR study to investigate how the optimisation of its side chain affects its *in vitro* activity, initially beginning with exploration of the Topliss scheme. The chemotypes of the other hits may also form the basis of similar investigation, and they may all form the basis of another round of virtual screening.

Chapter V described the use of molecular docking to rationalise the activity of compounds which inhibit malaria through inhibition of the bc_1 complex. Following the successful development of a suitable docking protocol, this was applied to rationalise the observed activity of a number of known *Pf*bc_1 inhibitors, including stigmatellin, atovaquone, antimycin A, HDQ, as well as several other quinolone containing compounds which had been previously been synthesised and tested at Liverpool. Inhibition of bc_1 has been attributed to binding at both the Q_o and Q_i sites of the complex, with Chapter V providing detailed discussion of the crucial interactions at either site, required for antimalarial activity. Methods were also developed to be able to distinguish between Q_o and Q_i inhibitors, as well as further validation sought for the 5 active hits from virtual screening and their potential as *Pf*bc_1 inhibitors. Combined, these studies show that molecular docking has many applications in modern drug discovery, and not only can it be used to explain a

compounds activity profile through consideration of observed interactions, but it has also proven useful in confirming the identity of novel antimalarial targets, and in rationalising observations with regard to emerging resistance. With work ongoing to design and validate a homology model of *Pf*bc$_1$, docking will help to drive forward ongoing efforts to further refine the molecules.

In Chapter VI a series of QSAR models were discussed that were generated for two datasets. The first dataset was that of a set of 45 4-aminoquinoline compounds that had been tested against both the NF54 and K1 strains of malaria. MLR, PLS and *k*NN machine learning methods were investigated, with the molecular descriptors contained within valid models interpreted with reference to the mechanistic mode of action of the 4-aminoquinoline compounds. Significant models were identified and shown to have strong predictive abilities for both strains of the parasite, and may ultimately prove useful in the future for the prediction of activity for similar 4-aminoquinoline compounds. The second dataset was that of a series of thiazolide containing compounds which had shown activity against HCV. After investigating several machine learning methods, SVM was found to give a significant model for the primary assay genotype 1B CC_{50} dataset, with this model able to predict the cell safety indices of the thiazolide derivatives. Combined, the 4-aminoquinoline and thiazolide studies illustrate the power and potential of QSAR methods, as both have led to predictive models which can be used to aid in drug design and safety respectively.

Finally, Chapter VII detailed the rational design of the novel pyrroloquinolone chemotype for antimalarial investigation. A total of 7 synthetic analogues were made using alkylation and Winterfeldt oxidation reactions, with compounds reporting activity values between 75 nM and 1.02 μM against the 3D7 malaria

parasite. A SAR investigation showed the chemotypes to have much potential, with molecular docking offering further support for possible Q_o binding and therefore *Pf*bc_1 inhibition for these compounds. Future work with regard to the pyrroloquinolone chemotype will involve improving the solubility of the motif, as well as investigating additional substitutions on the aromatic side chain to improve potency.

Experimental Chapter

Ligand Based Virtual Screening Methods

Fingerprint Similarity Searching Protocol

Fingerprint similarity searching protocol built and utilised within Pipeline Pilot Student Edition v6.1.[1] Structures manipulated using the SMILES chemical language.[2, 3] Twelve compounds active against *Pf*bc$_1$ tagged as reference structures. References used to screen the Zinc lead like library[4, 5] of compounds. Several molecular fingerprint methods used: ECFP_2; ECFP_4; ECFP_6; FCFP_2; FCFP_4; FCFP_6; MDLPublicKeys.[6, 7] Similarity between reference structures and those in the chemical library assessed using the Tanimoto coefficient.[8] Compounds reported as hits provided they fell within minimum and maximum similarity cut-off values of 0.70 and 0.99 respectively. All other settings left as standard. Reference structures and duplicate molecules removed. Hits reported in SMILES format with highest Tanimoto coefficient values recorded.

Turbo Similarity Searching Protocol

Turbo similarity searching protocol built and utilised within Pipeline Pilot Student Edition v6.1.[1] Structures manipulated using the SMILES chemical language.[2, 3] Twelve compounds active against *Pf*bc$_1$ tagged as reference structures. References used to screen the Zinc lead like library[4, 5] of compounds. Several molecular fingerprint methods used: ECFP_2; FCFP_2; MDLPublicKeys.[6, 7] Similarity between reference structures and those in the chemical library assessed using the Tanimoto coefficient.[8] Compounds reported as hits provided they fell within minimum and maximum similarity cut-off values of 0.80 and 0.99 respectively. All other settings left as standard. The hits from ZINC with the highest Tanimoto coefficient values capped at a maximum of 250 molecules. These 250 molecules are

used in a second iteration of similarity searching, using the same fingerprint method. Results from the two searches are merged, with reference structures and duplicate molecules removed. Hits reported in SMILES format with highest Tanimoto coefficient values recorded.

Bioisostere Substructure Searching Protocol

Quinolone core used as bioisostere query[9, 10] and reported using SMILES chemical language.[2, 3] BROOD version 1.1.2[11] used to identify bioisosteres of the query in the f5 and f50 fragment libraries. Searches performed four times using different searching methods: color; elect; struc; queryAnalog. Fragments ranked according to shape and chemistry. Hitlists merged in Pipeline Pilot Student Edition v6.1[1] and novel fragments identified according to their ring assemblies using the *'Find Novel Fragments'* component. Dataset filtered to include only neutral fragments with a ring count greater than one. Resulting bioisosteres used to perform a substructure search of the Zinc lead like library.[4, 5] When the number of hits for a particular bioisostere exceeds 200, a representative sample of 200 compounds is taken using the *'Diverse Molecules'* component with FCFP_4 fingerprints.[6, 7] Results merged. Hits reported in SMILES format. All other settings left at default.

Principal Component Analysis Protocol

PCA model for the twelve active compounds (tagged as reference structures) built using Pipeline Pilot Student Edition v6.1.[1] Structures manipulated using the SMILES chemical language.[2, 3] Physicochemical descriptors calculated for each compound: AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings. Descriptors used to calculate PCs, with PCA model developed to ensure greater than

75% of the variance in the data is explained. Model applied to the Zinc lead like library[4, 5] of compounds and mapped onto PC space. Molecules assessed according to their Euclidean distance[8] from one of the reference structures, with the 5,000 closest molecules selected and reported in SMILES format. All other settings left at default.

Naïve Bayesian Classification Protocol

Naïve Bayesian classification performed based on the structures of the twelve active and seven inactive compounds, qualitatively labelled 'yes' or 'no' respectively. Structures manipulated using the SDF file format. Physicochemical descriptors calculated using Pipeline Pilot Student Edition v6.1[1]: AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings; Num_Fragments. Bayesian model built within KNIME 2.3.3.[12] Data filtered to remove descriptors with low variance or which were highly correlated (correlation threshold 0.9). Descriptors normalised (Min-Max) and data partitioned as required. Additionally, SMOTE[13] could be performed at this stage. Classification model built using *'Naïve Bayes Learner'* component then applied to predict the class membership of external compounds with *'Naïve Bayes Predictor'*. Classification predictions reported in a confusion matrix. Further validation sought using 100 iterations of LOO cross validation. Validated Bayesian model applied to Zinc lead like library,[4, 5] with compounds defined as either active or inactive. Potentially active compounds filtered off and reported in an SDF file. All settings left at default unless otherwise stated.

Decision Tree Analysis Protocol

Decision tree analysis performed based on the structures of the twelve active and seven inactive compounds, qualitatively labelled 'yes' or 'no' respectively. Structures manipulated using the SDF file format. Physicochemical descriptors calculated using Pipeline Pilot Student Edition v6.1[1]: AlogP; Molecular_Weight; Num_H_Donors; Num_H_Acceptors; Num_RotatableBonds; Num_Atoms; Num_Rings; Num_AromaticRings; Num_Fragments. Decision tree built within KNIME 2.3.3.[12] Data filtered to remove descriptors with low variance or which were highly correlated (correlation threshold 0.9). Descriptors normalised (Min-Max) and data partitioned as required. Additionally, SMOTE[13] could be performed at this stage. Decision tree built using *'J48 Weka'*[14] node, then applied to predict the class membership of external compounds with *'Decision Tree Predictor'* component. Classification predictions reported in a confusion matrix. Further validation sought using 100 iterations of LOO cross validation. Validated decision tree model applied to Zinc lead like library,[4, 5] with compounds defined as either active or inactive. Potential actives filtered off and reported in an SDF file. All settings left at default unless otherwise stated.

Ligand Based Virtual Screening Scoring & Selection Methods

Consensus Scoring Protocol

Consensus scoring implemented in Pipeline Pilot Student Edition v6.1[1] using custom scripts and suitable components. See Chapter III for more details.

Compound Filtering Protocol

Compound filtering implemented in Pipeline Pilot Student Edition v6.1[1] using custom scripts and substructure filters. See Chapter III for more details.

Diverse Compound Selection Protocol

Diverse compound selection implemented in Pipeline Pilot Student Edition v6.1[1] using *'Diverse Molecules'* component and FCFP_4 fingerprints.[6, 7] All other settings left at default.

AGNES Clustering Protocol

AGNES[15] clustering performed in Pipeline Pilot Student Edition v6.1[1] utilising the R programming language.[16] Compounds clustered into required number of clusters using the *'R Cluster Agnes'* component according to their FCFP_4 fingerprints,[6, 7] with a sample molecule selected from each cluster using the *'Diverse Molecules'* component and FCFP_4 fingerprints. All other settings left at default.

CLARA Clustering Protocol

CLARA[15] clustering performed in Pipeline Pilot Student Edition v6.1[1] utilising the R programming language.[16] Compounds clustered into required number of clusters using the *'R Cluster Clara'* component according to their FCFP_4 fingerprints,[6, 7] and the central molecule from each selected. All other settings left at default.

Biological Assay Methods

Whole Cell Growth Inhibition Assay (3D7) Protocol

Compounds tested for 3D7 inhibition according to the standard procedure detailed below:[17-23]

P. falciparum (3D7 strain) cultures consisted of a 2% (v/v) suspension of O+ erythrocytes in RPMI-1640 medium (R 8758, glutamine and $NaHCO_3$) supplemented with 10% pooled human AB+ serum, 25 mM HEPES (pH 7.4), and 20 µM gentamicin sulphate. Cultures were grown under a gaseous headspace of 4% O_2, 3% CO_2 in N_2 at 37ºC. Parasite growth was synchronized by treatment with sorbitol. Drug susceptibilities were assessed by the measurement of fluorescence after the addition of SYBR Green I. Drug IC_{50} values were calculated from the log of the dose/response relationship, as fitted with Grafit software (Erithacus Software, Kent, UK). The results are given as the means of at least three separate experiments.

Complex I (NDH2) Bioassay Protocol

Compounds tested for NADH:decylubiquinone oxidoreductase inhibition according to the standard procedure detailed below:[23-25]

Recombinant *Pf*NDH2 activities were assayed in a reaction medium consisting of 50 nM potassium phosphate (pH 7.5), 2 mM EDTA, 200 µM NADH and 10 mM KCN. NADH:decylubiquinone oxidoreductase activity was initiated by the addition of 50 µM decylubiquinone. Decylubiquinone reduction was monitored at 283 nm (ε_{283} = 8.1 mM^{-1} cm^{1}) and 340 nm (ε_{340} = 6.22 mM^{-1} cm^{1}) in a Cary 4000 spectrophotometer. Percentage inhibition of NADH:decylubiquinone

oxidoreductase activity for drugs at a concentration of 28 µM determined using Excel.[26]

Bovine Complex III (bc_1) Bioassay Protocol

Compounds tested for cytochrome c reductase inhibition according to the standard procedure detailed below:[17, 23, 25]

Cytochrome *c* reductase activity measurements were assayed in 50 mM potassium phosphate (pH 7.5), 2 mM EDTA, 10 mM KCN, and 30 µM equine cytochrome *c* (Sigma) at room temperature. Cytochrome *c* reductase activity was initiated by the addition of decylubiquinol (50 µM). Reduction of cytochrome *c* was monitored in a Cary 4000 spectrophotometer at 550 versus 542 nm. Initial rates (computer-fitted as zero-order kinetics) were measured as a function of decylubiquinol concentration. Turnover rates of cytochrome *c* reduction were determined using $\varepsilon_{550\text{-}542} = 18.1\ mM^{-1}\ cm^{-1}$. Inhibitors of bc_1 activity were added without prior incubation. DMSO in the assays did not exceed 0.3 % (v/v). Percentage inhibition of cytochrome c reductase activity for drugs at a concentration of 56 µM determined using Excel.[26]

Molecular Docking Methods

Q_o Docking Protocol

- Protein-ligand molecular docking at the Q_o binding site performed using GOLD 5.0.1.[27]
- Wizard utilised to setup and performing docking calculations.
- Load appropriate file of reduced 3CX5 protein which includes SMA bound in the Q_o active site (pdb format).[28]
- Hydrogen atoms added to the protein.

- HOH7187 water molecule extracted for inclusion in docking calculation.[28-30] All other crystallographic water molecules removed.
- Tautomeric state of His181 altered such that the hydrogen atom forms a H-bond with the carbonyl group of SMA, and the lone pair of electrons on the nitrogen of His181 are directed towards ISP, in line with literature precedent.[30, 31]
- SMA ligand removed and used to define the Q_o binding site, together with all atoms around the ligand within 6Å.
- Ligand file/s loaded containing the compound/s to be docked (sdf format). As standard, for each ligand 10 GA runs are performed.
- If re-docking the native ligand, then load said ligand to enable RMSD comparison.
- Select the required fitness scoring function i.e. GOLDScore.[32, 33] If rescoring is required then select an additional scoring method i.e. ChemScore.[34-36]
- Option for early termination turned off.
- Default search efficiency used.
- All parameters left as standard, unless otherwise stated.
- Submit calculation and review results.

Qi Docking Protocol

- Protein-ligand molecular docking at the Q_i binding site performed using GOLD 5.0.1.[27]
- Wizard utilised to setup and performing docking calculations.
- Load appropriate file of reduced 3CX5 protein which includes ubiquinone bound in the Q_i active site (pdb format).[28]

- Hydrogen atoms added to the protein.
- All crystallographic water molecules removed.
- Ubiquinone ligand removed and used to define the Q_i binding site, together with all atoms around the ligand within 6Å.
- Ligand file/s loaded containing the compound/s to be docked (sdf format). As standard, for each ligand 10 GA runs are performed.
- Select the required fitness scoring function i.e. GOLDScore.[32, 33] If rescoring is required then select an additional scoring method i.e. ChemScore.[34-36]
- Option for early termination turned off.
- Default search efficiency used.
- All parameters left as standard, unless otherwise stated.
- Submit calculation and review results.

Additional Computational Methods

Energy Minimisation Protocol

Spartan '08 was used to perform energy minimisation calculations.[37] The required molecule for calculation was built using the construction tools within Spartan. An equilibrium geometry calculation was then performed at ground state, using the molecular mechanics, MMFF level of theory.[38] The total charge was adjusted accordingly and all other settings left at default.

Conformer Distribution Protocol

Spartan '08 was used to perform conformer distribution calculations.[37] The required molecule for calculation was built using the construction tools within Spartan. A conformer distribution calculation was then performed at ground state, using the

molecular mechanics, MMFF level of theory.[38] All other settings left at default. One hundred conformations were calculated, with a spreadsheet used to record the energy and Boltzmann distribution values for each conformer.

Conformer Library Protocol

Spartan '08 was used to perform conformer library calculations.[37] The required molecule for calculation was built using the construction tools within Spartan. A conformer library calculation was then performed, in which the lowest energy conformer of a molecule is replaced with a set of conformers spanning all shapes accessible to the molecule, with no regard for its energy.[39] All settings left at default.

Similarity Analysis Protocol

Spartan '08 was used to perform similarity analysis calculations.[37] The pharmacophore with which to map additional molecules against is loaded. A similarity analysis calculation can then be performed, mapping each compound in a specified conformer library to the pharmacophore. Similarity can be assessed using the CFDs of the pharmacophore, with each compound assigned a score between 0 and 1.[39] All other settings left at default.

Substructure Searching Protocol

Substructure search protocol built and utilised within Pipeline Pilot Student Edition v6.1.[1] Structures manipulated in their SDF format. Library of compounds searched for a query structure using the *'Substructure Filter from File'* component. Results written in SDF format, all other settings left at default.

Fingerprint Similarity Search Protocol

Fingerprint similarity search protocol built and utilised within Pipeline Pilot Student Edition v6.1.[1] Structures manipulated in their SDF format. Query structure used to screen a library of compounds using FCFP_4 molecular fingerprints.[6, 7] Similarity between query and library compounds assessed using the Tanimoto coefficient.[8] Compounds filtered according to required level of similarity. Results written in SDF format, all other settings left at default.

QSAR Experimental Procedures

QSAR Protocol 1 *MLR; Training set*

The following procedure was used to develop QSAR models using MLR when all of the molecules formed part of a training set. Calculation of internal validation statistics are also discussed:

- Load the data into PHAKISO[40] as a training set.
- Use Autoscale to normalise the descriptors.
- Perform objective descriptor selection:
    - General descriptor selection. (Removes descriptors with the same value for 80% of the training set and those with missing values.)
    - CORCHOP[41] descriptor selection. (Removes descriptors with very highly correlated R values (0.99), with a maximum allowed correlation of 0.75, and a maximum allowed kurtosis of 100.)
- Perform subjective descriptor selection using the required selection method: Forward selection; Backward elimination; Stepwise Regression; Genetic algorithm; GALib. Adjusted coefficient of determination used as the error

measurement. Max/min number of variables altered accordingly to produce favourable molecule/descriptor ratios.

- Train the data using MLR machine learning.[8]
- Predict the internal statistics. Process automated through PHAKISO to calculate statistical parameters such as the correlation coefficient, coefficient of determination, F-statistic etc. Additionally the cross validation parameter q^2 can be calculated, as can the N fold cross validation[40] and Bootstrapping[42] statistics.
- Calculate the t-statistics for the descriptors using the data analysis, regression tools in Excel.[26, 43]
- Tabulate the results in a summary file.

QSAR Protocol 2 *MLR; Training and test set*

The following procedure was used to divide a dataset into a training and test set. QSAR models were then developed using MLR. Calculation of internal and external validation statistics are also discussed:

- Load the data into PHAKISO[40] as a dataset.
- Divide the data into a training and test set using the most appropriate method:
    - Sphere exclusion algorithm. (All settings left at default unless stated.)
    - CADEX. (All settings left at default unless stated.)
    - Activity binning. (Performed in Excel.)[26]
- Use Autoscale to normalise the descriptors.
- Perform objective descriptor selection:

  - General descriptor selection. (Removes descriptors with the same value for 80% of the training set and those with missing values.)
  - CORCHOP[41] descriptor selection. (Removes descriptors with very highly correlated R values (0.99), with a maximum allowed correlation of 0.75, and a maximum allowed kurtosis of 100.)

- Perform subjective descriptor selection using the required selection method: Forward selection; Backward elimination; Stepwise Regression; Genetic algorithm; GALib. Adjusted coefficient of determination used as the error measurement. Max/min number of variables altered accordingly to produce favourable molecule/descriptor ratios.
- Train the data using MLR machine learning.[8]
- Predict the internal statistics. Process automated through PHAKISO to calculate statistical parameters such as the correlation coefficient, coefficient of determination, F-statistic etc. Additionally the cross validation parameter q^2 can be calculated, as can the N fold cross validation[40] and Bootstrapping[42] statistics.
- Calculate the t-statistics for the descriptors using the data analysis, regression tools in Excel.[26, 43]
- Apply internally validated model to the test set.
- Predict the external statistics. Process automated through PHAKISO to calculate statistical parameters such as the correlation coefficient, coefficient of determination, F-statistic etc.
- In Excel plot the predicted vs. actual and actual vs. predicted activity values for the training and test sets.
- Calculate the Tropsha parameters using the graphs.[44, 45]

- Tabulate the results in a summary file.

QSAR Protocol 3 *PLS; Training and test set*

The following procedure was used to divide a dataset into a training and test set. QSAR models were then developed using PLS. Calculation of internal and external validation statistics are also discussed:

- Load the data into PHAKISO[40] as a dataset.
- Divide the data into a training and test set using the most appropriate method:
    - Sphere exclusion algorithm. (All settings left at default unless stated.)
    - CADEX. (All settings left at default unless stated.)
    - Activity binning. (Performed in Excel.)[26]
- Use Autoscale to normalise the descriptors.
- Find the optimum number of components to explain the data. Usually when the combined components have a q^2 > 0.5, with the last component increasing the explained variance by more than 5%. (Use the single parameter trainer with PLS and adjusted coefficient of determination.)
- Set the number of components and train the data using the PLS machine learning method.[46]
- Predict the internal statistics. Process automated through PHAKISO to calculate statistical parameters such as the correlation coefficient, coefficient of determination, *F*-statistic etc. Additionally the cross validation parameter q^2 can be calculated, as can the N fold cross validation[40] and Bootstrapping[42] statistics.
- Apply internally validated model to the test set.

- Predict the external statistics. Process automated through PHAKISO to calculate statistical parameters such as the correlation coefficient, coefficient of determination, F-statistic etc.
- In Excel plot the predicted vs. actual and actual vs. predicted activity values for the training and test sets.
- Calculate the Tropsha parameters using the graphs.[44, 45]
- Tabulate the results in a summary file.

QSAR Protocol 4 *kNN*

The following procedure was used to develop models with the *k*NN machine learning method:

- Construct two text files, one which contains the molecule names and associated descriptors, the other the molecule names with their activities.
- Use unsupervised forward selection (UFS) to reduce the number of descriptors. UFS was designed for use in the development of QSARs as a data reduction algorithm that selects from a data matrix, a maximal linearly independent set of columns with a minimal amount of multiple correlation i.e. it removes highly correlated descriptors.[47]
- Normalize the data.
- Run the *k*NN calculation through *Cygwin* using codes which were developed by Dr N. Berry at Liverpool University from executable codes provided by Prof. A. Tropsha at UNC. *Cygwin* is a unix-like environment and command-line interface for Microsoft Windows.[48] Conditions can be specified as to the splitting of the data into training and test sets, as well as the number of models per split, and the maximum and minimum descriptor range.

- The codes construct all possible models within the specified conditions, and randomise the data such that all statistically good models are reported at varying significance levels, and grouped accordingly.

QSAR Protocol 5 *SVM; Training and test set*

The following procedure was used to divide a dataset into a training and test set. QSAR models were then developed using SVM. Calculation of internal and external validation statistics are also discussed:

- In PHAKISO[40] use Autoscale to normalise the descriptors.
- Perform objective descriptor selection:
    - General descriptor selection. (Removes descriptors with the same value for 80% of the training set and those with missing values.)
    - CORCHOP[41] descriptor selection. (Removes descriptors with very highly correlated R values (0.99), with a maximum allowed correlation of 0.75, and a maximum allowed kurtosis of 100.)
- Divide the data into a training and test set using the most appropriate method:
    - CADEX. (All settings left at default unless stated.)
    - Activity binning. (Performed in Excel.)[26]
- SVM protocol built using KNIME.[12]
- Training set read in.
- Low variance and linear correlation filters performed, both at default settings.
- SVM regression models built using the radial basis function (RBF) kernel via the GridSearch node. The grid search algorithm found the SVM model with RBF kernel parameters ε and γ that gave the highest correlation coefficient in 10-fold cross-validation. The ε parameter was initially altered in a range of

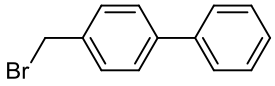
2^{-10} to 2^{1}, and the γ parameter was from 2^{-15} to 2^{3}, in line with recommendations in the literature.[49] Both parameters were altered as powers of 2, and the grid search was allowed to extend three times. The C parameter for the RBF was manually scanned from 1 to 1000 in factors of 10 initially, and then in half order of magnitude once a coarse optimum region of parameter space was identified.

- Internal and external validation statistics calculated using Excel by plotting the predicted vs. actual and actual vs. predicted activity values for the training and test sets.
- Calculate the Tropsha parameters using the graphs.[44, 45]
- Tabulate the results in a summary file.

Chemical Synthesis

All reactions were carried out in dry conditions under a nitrogen atmosphere unless otherwise stated. Melting points were determined with a Gallenkamp apparatus and are uncorrected. Elemental analysis was performed by the microanalysis service at the University of Liverpool. Mass spectra were collected using a Fisons TRIO 1000 by the mass spec lab at the University of Liverpool. Infrared spectra were recorded on a Perkin Elmer 4100-typeA FTIR spectrometer. ^{1}H-NMR spectra were recorded on a Bruker AMX 400 (400MHz) spectrometer, as were ^{13}C-NMR spectra in solutions of $CDCl_3$ and MeOD. The chemical shifts are in parts per million (ppm), with tetramethylsilane as the internal reference and the coupling constants in hertz (Hz). TLC was performed on silica plates, and columns were run on silica gel specifically for flash chromatography. Reagents were purchased from Sigma-Aldrich.

Wohl-Ziegler Bromination Reaction

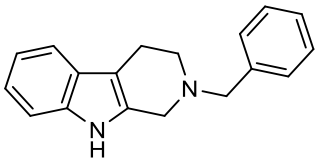


4-(bromomethyl)-1,1'-biphenyl (7).

4-methyl-1,1'-biphenyl (0.3146 g, 1.870 mmol) was dissolved in acetonitrile (40 mL) and the system flushed with nitrogen. AIBN (0.0976 g, 0.5944 mmol) and NBS (0.9521 g, 5.350 mmol) were then added to the solution and left to stir at reflux until all of the biphenyl was consumed. The reaction was then cooled to room temperature and the solid brown residue filtered off. The orange filtrate was then concentrated to give the crude product as an orange solid. The product was then purified via flash column chromatography (silica gel, 3:97 EtOAc/Hexane) to give the pure product as an orange solid (0.4095 g, 88.61%).; mp 81^{o}C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 4.51 (s, 2H), 7.31-7.61 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100 MHz): 33.9, 127.9, 128.4, 129.1, 129.4, 130.0, 137.2, 140.9, 141.8; IR (neat cm^{-1}) 2978, 2357, 1485; MS (m/z) 246 [M-H], 182 [M-C_6H_5], 168 [M-Br]; Anal. Calcd for $C_{13}H_{11}Br$: C, 63.18%; H, 4.49%. Found: C, 56.26%; H, 3.86%.

Experimental data in agreement with literature precedent.[50]

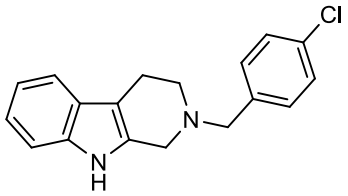
Alkylation Reactions



2-benzyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (8). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (0.5092 g, 2.957 mmol) was dissolved in anhydrous THF (40

mL) at 0°C under inert conditions. Triethylamine (0.82 mL, 5.883 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. (Bromomethyl)benzene **(1)** (0.42 mL, 3.531 mmol) was then added and left to stir until all of the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole was consumed. Once the solvents were removed, the crude product was purified via flash column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as a pale yellow solid (0.5096 g, 65.73%).; mp 138°C; ^{1}H NMR δ ppm ($CDCl_3$, 400MHZ): 2.71 (t, *J*=5.3 Hz, 2H), 2.79 (t, *J*=5.3 Hz, 2H), 3.42 (s, 2H), 3.63 (s, 2H), 6.96-7.38 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.6, 50.5, 51.3, 62.4, 108.7, 111.2, 118.4, 119.7, 121.7, 127.7, 127.7, 128.9, 129.6, 132.3, 136.4, 138.8; IR (neat cm^{-1}) 2978, 2360,1454; MS (m/z) 263 $[M+H]^+$; HRMS (CI) calcd for $C_{18}H_{19}N_2$ (MH^+) requires 263.1544, found 263.1548. Anal. Calcd for $C_{18}H_{18}N_2$: C, 82.41%; H, 6.92%; N, 10.68%. Found: C, 82.36%; H, 6.96%; N, 10.60%.

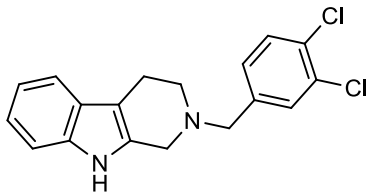
Experimental data in agreement with literature precedent.[51, 52]



2-(4-chlorobenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (9). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (1.0 g, 5.806 mmol) was dissolved in anhydrous THF (80 mL) at 0°C under inert conditions. Triethylamine (1.62 mL, 11.612 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 4-chlorobenzyl bromide **(2)** (1.45 g, 7.127 mmol) was then added and left to stir until all of the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole was consumed. Once the solvents were removed, the crude product was purified via flash column

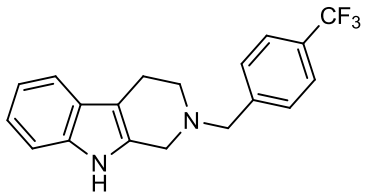
chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as a yellow solid (1.2969 g, 75.33%).; mp 186^{o}C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 2.83 (t, *J*=5.3 Hz, 2H), 2.90 (t, *J*=5.3 Hz), 3.66 (s, 2H), 3.74 (s, 2H), 7.06-7.48 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 22.5, 41.8, 55.8, 61.5, 108.6, 111.1, 118.3, 119.6, 121.7, 127.8, 128.9, 130.9, 133.4, 133.5, 137.4, 137.6; IR (neat cm^{-1}) 3132, 3059, 2924, 2839; MS (m/z) 297 $[M+H]^+$; HRMS (CI) calcd for $C_{18}H_{18}N_2Cl$ (MH^+) requires 297.11585, found 297.11559. Anal. Calcd for $C_{18}H_{17}N_2Cl$: C, 72.83%; H, 5.77%; N, 9.44%. Found: C, 72.63%; H, 5.82%; N, 9.39%.

Experimental data in agreement with literature precedent.[53]

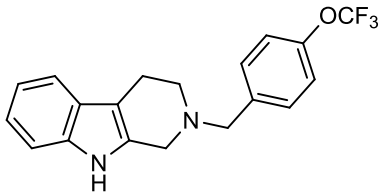


2-(3,4-dichlorobenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (10). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (1.0 g, 5.806 mmol) was dissolved in anhydrous THF (80 mL) at 0^{o}C under inert conditions. Triethylamine (1.62 mL, 11.612 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 3,4-dichlorobenzyl bromide **(3)** (1.1 mL, 7.565 mmol) was then added and left to stir until all the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b] was consumed. Once the solvents were removed, the crude product was purified via flash column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as a pale orange solid (1.5258 g, 79.39%).; mp 158^{o}C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 2.83 (t, *J*=5.2Hz, 2H), 2.91 (t, *J*=5.2Hz, 2H), 3.67 (s, 2H), 3.72 (s, 2H), 7.07-7.53 (m, 7H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.5, 50.6, 51.3, 61.1, 108.9, 111.1, 118.4, 119.9, 121.9, 128.6, 130.7, 131.1, 131.9, 133.2, 133.8, 135.9, 136.4, 139.5; IR (neat

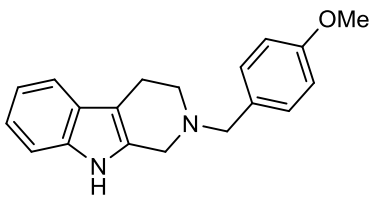
cm^{-1}) 3059, 2958, 2916, 2839; MS (m/z) 331 $[M+H]^+$; HRMS (CI) calcd for $C_{18}H_{17}N_2Cl_2$ (MH^+) requires 331.07688, found 331.07736. Anal. Calcd for $C_{18}H_{16}N_2Cl_2$: C, 65.25%; H, 4.87%; N, 8.46%. Found: C, 65.07%; H, 4.84%; N, 8.34%.



2-(4-(trifluoromethyl)benzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (11). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (1.0 g, 5.806 mmol) was dissolved in anhydrous THF (80 mL) at 0^oC under inert conditions. Triethylamine (1.62 mL, 11.612 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 4-(trifluoromethyl)benzyl bromide **(4)** (1.67 g, 6.986 mmol) was then added and left to stir until all of the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole was consumed. Once the solvents were removed, the crude product was purified via flash column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as a pale yellow solid (1.6377 g, 85.50%).; mp 167^oC; ^1H NMR δ ppm ($CDCl_3$, 400MHZ): 2.82 (t, *J*=5.3 Hz, 2H), 2.89 (t, *J*=5.3 Hz, 2H), 3.61 (s, 2H), 3.79 (s, 2H), 7.06-7.60 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.6, 50.6, 51.3, 61.7, 108.6, 111.1, 118.3, 119.8, 121.8, 123.3, 125.6, 126.0, 127.6, 129.6, 132.0, 136.5, 143.2; IR (neat cm^{-1}) 3143, 3062, 2941, 2833; MS (m/z) 331 $[M+H]^+$; HRMS (CI) calcd for $C_{19}H_{18}N_2F_3$ (MH^+) requires 331.14221, found 331.14199. Anal. Calcd for $C_{19}H_{17}N_2F_3$: C, 69.08%; H, 5.19%; N, 8.48%. Found: C, 69.00%; H, 5.22%; N, 8.45%.

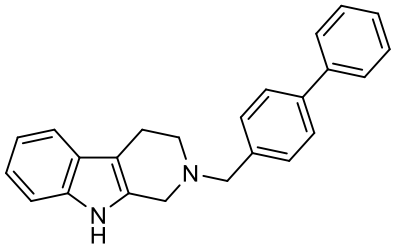


2-(4-(trifluoromethoxy)benzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (12). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (1.0 g, 5.806 mmol) was dissolved in anhydrous THF (80 mL) at 0^{o}C under inert conditions. Triethylamine (1.62 mL, 11.62 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 4-(trifluoromethoxy)benzyl bromide **(5)** (1.0 mL, 6.250 mmol) was then added and left to stir until all the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b] was consumed. Once the solvents were removed the crude product was purified via flash column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as an orange solid (1.7342 g, 86.24%).; mp 116^{o}C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 2.83 (t, *J*=5.3Hz, 2H), 2.91 (t, *J*=5.3Hz, 2H), 3.66 (s, 2H), 3.77 (s, 2H), 7.07-7.49 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.6, 50.6, 51.3, 61.4, 100.0, 108.8, 111.1, 118.4, 119.3, 121.3, 121.8, 127.6, 130.7, 132.0, 136.4, 137.7, 148.8; IR (neat cm^{-1}) 3180, 3100, 2825, 1508, 1269; MS (m/z) 347 $[M+H]^{+}$; HRMS (CI) calcd for $C_{19}H_{18}N_2OF_3$ (MH^{+}) requires 347.13712, found 347.13701. Anal. Calcd for $C_{19}H_{17}N_2OF_3$: C, 65.89%; H, 4.95%; N, 8.09%. Found: C, 65.25%; H, 4.89%; N, 7.94%.



2-(4-methoxybenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (13). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (0.5270 g, 3.060 mmol) was dissolved in

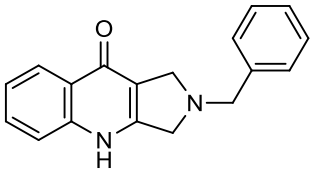
anhydrous THF (40 mL) at 0°C under inert conditions. Triethylamine (0.85 mL, 6.098 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 1-(bromomethyl)-4-methoxybenzene **(6)** (0.51 mL, 3.643 mmol) was then added and left to stir until all the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b] was consumed. Once the solvents were removed the crude product was purified via flash column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as an orange solid (0.4719 g, 52.75%).; mp 131°C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 2.71 (t, *J*=5.4Hz, 2H), 2.78 (t, *J*=5.4Hz, 2H), 3.41 (s, 2H), 3.67 (s, 2H), 6.73-7.37 (m, 8H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.6, 50.5, 51.3, 55.7, 62.3, 108.6, 111.2, 114.8, 118.4, 119.7, 122.0, 127.7, 129.8, 132.3, 136.5, 140.5, 160.2; IR (neat cm^{-1}) 2939, 1971, 1597, 1462; MS (m/z) 293 $[M+H]^+$; HRMS (CI) calcd for $C_{19}H_{21}N_2O$ (MH^+) requires 293.1647, found 293.1654. Anal. Calcd for $C_{19}H_{20}N_2O$: C, 78.05%; H, 6.89%; N, 9.58%. Found: C, 77.21%; H, 6.77%; N, 9.35%.



2-([1,1'-biphenyl]-4-ylmethyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole (14). 1,2,3,4-tetrahydro-9H-pyrido[3,4-b]indole (0.2904 g, 1.686 mmol) was dissolved in anhydrous THF (30 mL) at 0°C under inert conditions. Triethylamine (0.47 mL, 3.372 mmol) was added dropwise to the solution and left to warm to room temperature for 1 hour. 4-(bromomethyl)-1,1'-biphenyl **(7)** (0.5 g, 3.643 mmol) was then added and left to stir until all the 1,2,3,4-tetrahydro-9H-pyrido[3,4-b] was consumed. Once the solvents were removed the crude product was purified via flash

column chromatography (silica gel, 40:60 EtOAc/Hexane) to give the pure product as a pale yellow solid (0.2970 g, 45.80%).; mp 160°C; ^{1}H NMR δ ppm ($CDCl_3$, 400 MHz): 2.84 (t, *J*=5.4Hz, 2H), 2.94 (t, *J*=5.4Hz, 2H), 3.66 (s, 2H), 3.80 (s, 2H), 7.06-7.62 (m, 13H); ^{13}C NMR δ ppm ($CDCl_3$, 100MHz): 21.6, 50.6, 51.4, 62.1, 108.8, 111.1, 118.4, 119.8, 121.7, 127.5, 127.5, 127.7, 127.7, 129.2, 130.0, 132.3, 136.4, 137.9, 140.6, 141.3; IR (neat cm^{-1}) 3394, 2978, 1450; MS (m/z) 339 $[M+H]^+$; HRMS (CI) calcd for $C_{24}H_{23}N_2$ (MH^+) requires 339.1848, found 339.1861. Anal. Calcd for $C_{24}H_{22}N_2$: C, 85.17%; H, 6.55%; N, 8.28%. Found: C, 84.68%; H, 6.56%; N, 8.12%.

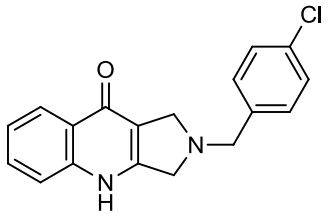
Winterfeldt Oxidation Reactions



2-benzyl-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (15). 2-benzyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(8)** (0.2972 g, 1.133 mmol) and potassium tert-butoxide (0.1290 g, 1.150 mmol) were dissolved in DMF (10 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (75 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a yellow solid (0.1016 g, 32.46%).; mp 222°C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.65 (s, 2H), 4.71 (s, 2H), 4.86 (s, 2H), 7.46-8.29 (m, 9H); ^{13}C NMR δ ppm

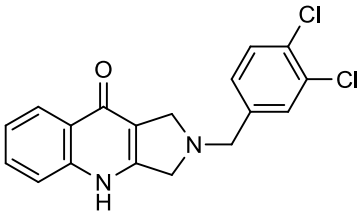
(MeOD + trifluoroacetic acid, 100MHz): 57.8, 57.9, 60.4, 113.0, 120.2, 126.4, 126.6, 126.9, 131.1, 131.8, 132.0, 134.5, 138.0, 142.3, 147.5, 175.5; IR (neat cm^{-1}) 2804, 2357, 1570, 1504; MS (m/z) 277 $[M+H]^+$; HRMS (CI) calcd for $C_{18}H_{17}N_2O$ (MH^+) requires 277.1341, found 277.1341. Anal. Calcd for $C_{18}H_{16}N_2O$: C, 78.24%; H, 5.84%; N, 10.14%. Found: C, 74.83%; H, 5.61%; N, 9.59%.

Experimental data in agreement with literature precedent.[54]

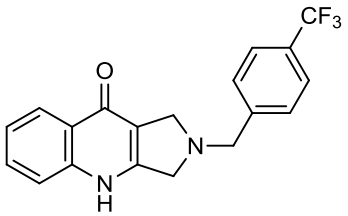


2-(4-chlorobenzyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (16). 2-(4-chlorobenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(9)** (0.5991 g, 2.019 mmol) and potassium tert-butoxide (0.2235 g, 1.992 mmol) were dissolved in DMF (15 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (150 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a yellow solid (0.4961 g, 79.08%).; mp 251^oC; ^1H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.64 (s, 2H), 4.70 (s, 2H), 4.87 (s, 2H), 7.45-8.28 (m, 8H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 57.8, 57.9, 59.5, 113.0, 120.3, 126.4, 126.5, 126.9, 130.6, 131.2, 133.7, 134.5, 138.1, 142.2, 147.5, 175.4; IR (neat cm^{-1}) 3064, 2926, 2783, 2364, 1898; MS (m/z) 311 $[M+H]^+$, 185 $[M-C_7H_6Cl]^+$; HRMS (CI) calcd for $C_{18}H_{16}ClN_2O$ (MH^+) requires 311.09512,

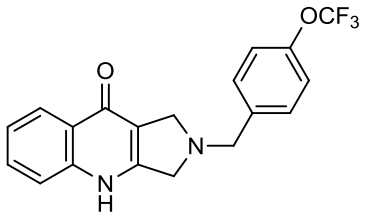
found 311.09378. Anal. Calcd for $C_{18}H_{15}ClN_2O$: C, 69.57%; H, 4.86%; N, 9.01%. Found: C, 69.16%; H, 4.92%; N, 8.84%.



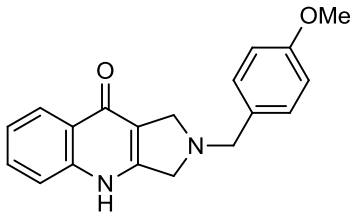
2-(3,4-dichlorobenzyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (17). 2-(3,4-dichlorobenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(10)** (0.5610 g, 1.694 mmol) and potassium tert-butoxide (213.60 mg, 1.903 mmol) were dissolved in DMF (15 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (150 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a yellow solid (0.2785 g, 47.63 %).; mp 244^{o}C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.66 (s, 2H), 4.70 (s, 2H), 4.89 (s, 2H), 7.46-8.29 (m, 7H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 57.9, 58.0, 58.9, 122.9, 120.3, 126.4, 126.5, 126.8, 132.0, 132.4, 133.2, 134.1, 134.5, 134.9, 135.7, 142.2, 147.5, 175.8; IR (neat cm^{-1}) 3080, 2929, 2810, 2349, 1845; MS (m/z) 345 $[M+H]^+$, 185 $[M-C_7H_5Cl_2]^+$; HRMS (CI) calcd for $C_{18}H_{15}Cl_2N_2O$ (MH^+) requires 345.05614, found 345.05507. Anal. Calcd for $C_{18}H_{14}Cl_2N_2O$: C, 62.63%; H, 4.09%; N, 8.11%. Found: C, 61.98%; H, 3.99%; N, 7.98%.



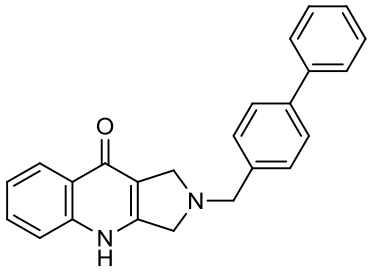
2-(4-(trifluoromethyl)benzyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (18). 2-(4-(trifluoromethyl)benzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(11)** (0.5128 g, 1.552 mmol) and potassium tert-butoxide (0.1754 g, 1.563 mmol) were dissolved in DMF (12 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (150 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a pale yellow solid (0.1321 g, 24.72%).; mp 265^{o}C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.67 (s, 2H), 4.81 (s, 2H), 4.90 (s, 2H), 7.46-8.29 (m, 8H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 58.0, 58.2, 59.5, 113.0, 117.9, 120.3, 126.4, 126.5, 126.9, 127.8, 127.9, 132.8, 134.5, 136.2, 142.2, 147.5, 175.4; IR (neat cm^{-1}) 3066, 2931, 2814, 2334, 1895; MS (m/z) 345 $[M+H]^{+}$, 185 $[M-C_8H_6F_3]^{+}$; HRMS (CI) calcd for $C_{19}H_{16}N_2OF_3$ (MH^{+}) requires 345.12147, found 345.12096. Anal. Calcd for $C_{19}H_{15}N_2OF_3$: C, 66.23%; H, 4.39%; N, 8.14%. Found: C, 63.70%; H, 4.38%; N, 6.81%.



2-(4-(trifluoromethoxy)benzyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (19). 2-(4-(trifluoromethoxy)benzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(12)** (0.6146 g, 1.775 mmol) and potassium tert-butoxide (0.2200 g, 1.960 mmol) were dissolved in DMF (16 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (150 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a yellow solid (0.2015 g, 31.51%).; mp 257^{o}C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.68 (s, 2H), 4.76 (s, 2H), 4.90 (s, 2H), 7.46-8.28 (m, 8H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 57.9, 58.0, 59.4, 113.0, 120.3, 123.4, 126.4, 126.5, 126.9, 130.9, 134.1, 134.5, 142.3, 147.5, 160.5, 160.9, 175.4; IR (neat cm^{-1}) 2812, 2337, 1570, 1512; MS (m/z) 361 $[M+H]^{+}$; HRMS (CI) calcd for $C_{19}H_{16}N_2O_2F_3$ (MH^{+}) requires 361.1147, found 361.1147. Anal. Calcd for $C_{19}H_{15}N_2O_2F_3$: C, 63.33%; H, 4.20%; N, 7.77%. Found: C, 63.23%; H, 4.24%; N, 7.72%.



2-(4-methoxybenzyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (20). 2-(4-methoxybenzyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(13)** (0.3007 g, 1.029 mmol) and potassium tert-butoxide (0.1170 g, 1.043 mmol) were dissolved in DMF (10 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (75 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a pale yellow solid (0.0749 g, 23.77%).; mp 216^{o}C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.65 (s, 2H), 4.67 (s, 2H), 4.85 (s, 2H), 7.11-8.29 (m, 8H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 56.4, 57.9, 60.4, 113.0, 115.7, 117.3, 117.4, 120.2, 123.8, 126.4, 126.6, 132.3, 133.1, 134.5, 142.3, 147.4, 175.4; IR (neat cm^{-1}) 3070, 2850, 2890, 1570, 1516; MS (m/z) 307 $[M+H]^{+}$; HRMS (CI) calcd for $C_{19}H_{19}N_2O_2$ (MH^{+}) requires 307.1444, found 307.1447. Anal. Calcd for $C_{19}H_{18}N_2O_2$: C, 74.49%; H, 5.92%; N, 9.14%. Found: C, 74.41%; H, 6.01%; N, 9.12%.



2-([1,1'-biphenyl]-4-ylmethyl)-2,3-dihydro-1H-pyrrolo[3,4-b]quinolin-9(4H)-one (21). 2-([1,1'-biphenyl]-4-ylmethyl)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole **(14)** (0.2560 g, 0.7564 mmol) and potassium tert-butoxide (0.0908 g, 0.8091 mmol) were dissolved in DMF (10 mL). A steady stream of oxygen was bubbled through the solution and the reaction was monitored via the consumption of the indole. Water (75 mL) was added and the solution was neutralised using 1M HCl. The precipitate was filtered off and an extraction was carried out on the filtrate using ethyl acetate. The solvents were removed and the two crops of solid combined. DCM was added to the collected solid and following vigorous shaking, the mixture was filtered to give the product as a yellow solid (0.1088 g, 40.81%).; mp 245^{o}C; ^{1}H NMR δ ppm (MeOD + trifluoroacetic acid, 400 MHz): 4.68 (s, 2H), 4.75 (s, 2H), 4.89 (s, 2H), 7.39-8.28 (m, 13H); ^{13}C NMR δ ppm (MeOD + trifluoroacetic acid, 100MHz): 57.4, 57.5, 59.7, 112.6, 119.8, 126.0, 126.2, 126.6, 128.1, 129.1, 130.1, 130.2, 132.1, 134.1, 141.1, 141.9, 144.6, 147.0, 175.1; IR (neat cm^{-1}) 2981, 1747, 1570, 1508; MS (m/z) 353 $[M+H]^{+}$; HRMS (CI) calcd for $C_{24}H_{21}N_2O$ (MH^{+}) requires 353.1646, found 353.1654. Anal. Calcd for $C_{24}H_{20}N_2O$: C, 81.79%; H, 5.72%; N, 7.95%. Found: C, 77.85%; H, 5.70%; N, 7.25%.

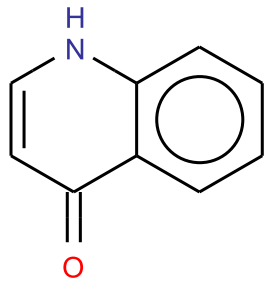
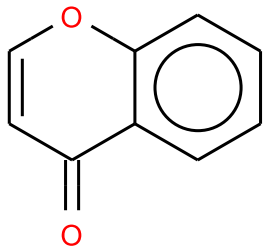
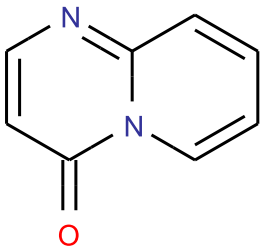
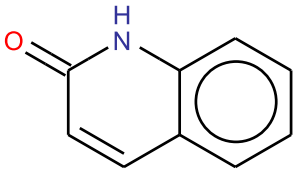
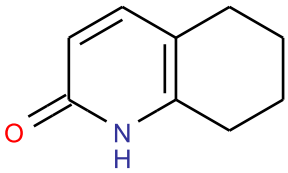
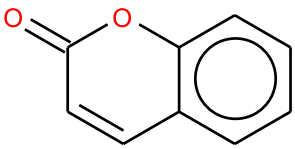
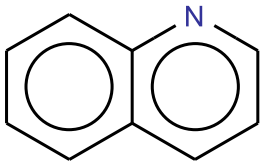
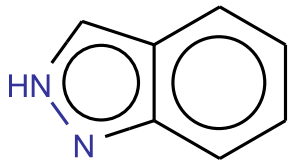
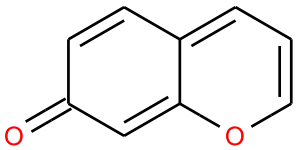
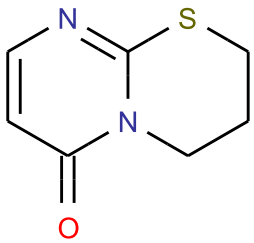
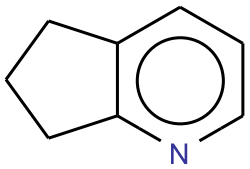
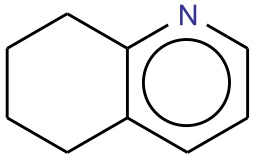
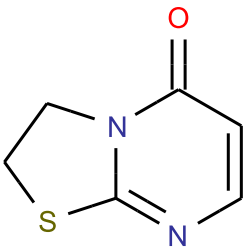
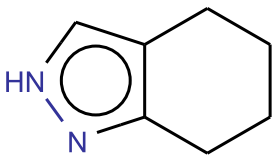
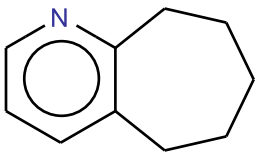
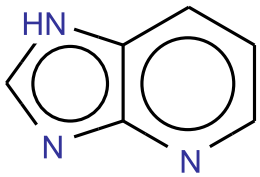
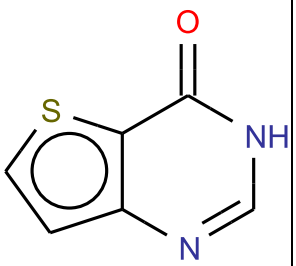
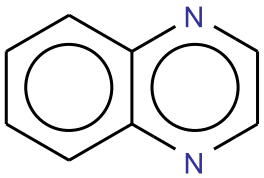
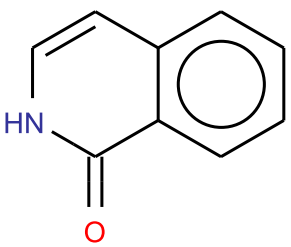
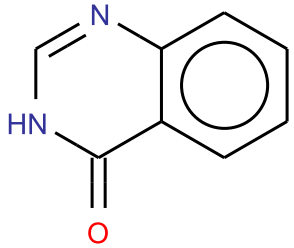
References

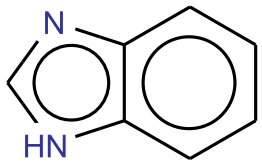
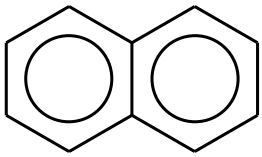
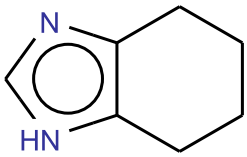
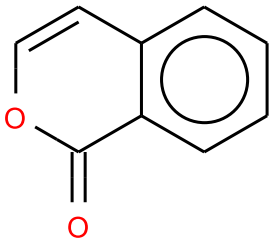
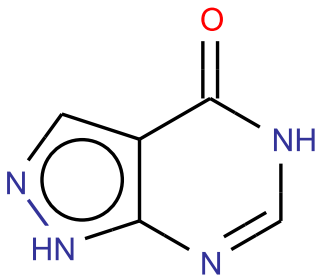
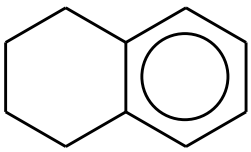
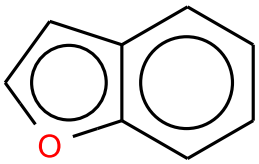
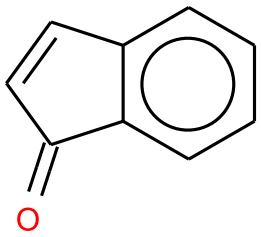
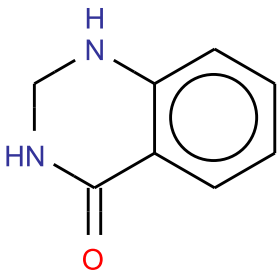
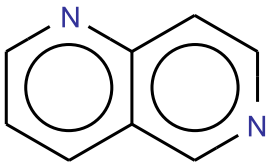
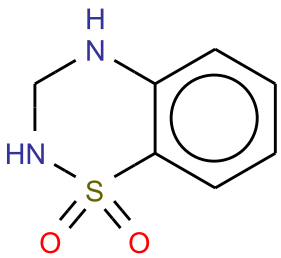
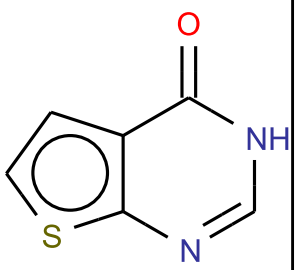
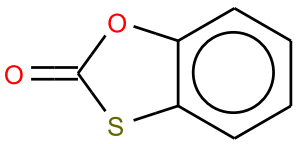
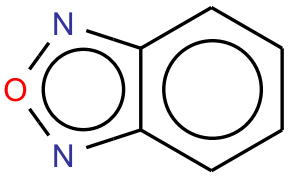
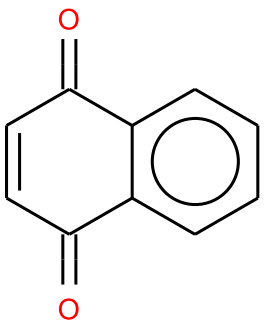
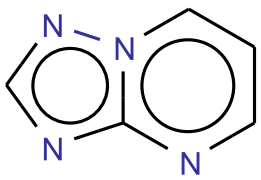
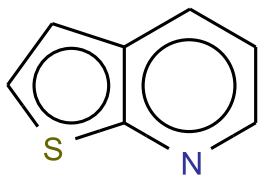
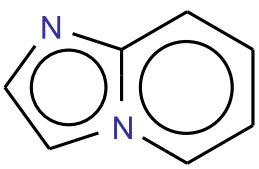
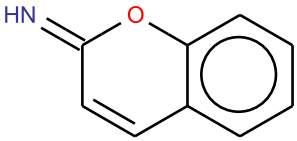
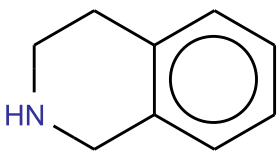
1. SciTegic, *Pipeline Pilot Student Edition v6.1*, Accelrys, Inc, San Diego, CA, 2007.
2. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
3. D. Weininger, A. Weininger and J. L. Weininger, *Journal of Chemical Information and Computer Sciences*, 1989, **29**, 97-101.
4. J. J. Irwin and B. K. Shoichet, *Journal of Chemical Information and Modeling*, 2005, **45**, 177-182.
5. S. J. Teague, A. M. Davis, P. D. Leeson and T. Oprea, *Angew. Chem.-Int. Edit.*, 1999, **38**, 3743-3748.
6. D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742-754.
7. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *Journal of Chemical Information and Computer Sciences*, 2002, **42**, 1273-1280.
8. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Springer, 2007.
9. A. M. Wassermann and J. Bajorath, *Future Medicinal Chemistry*, 2011, **3**, 425-436.
10. K. Kim, J. Kang, S. Kim, S. Choi, S. Lim, C. Im and C. Yim, *Archives of Pharmacal Research*, 2007, **30**, 570-580.
11. OpenEye, *BROOD version 1.1.2; http://www.eyesopen.com/brood*, Accessed October 2011.
12. www.knime.org, *KNIME v2.3.3*, 2003-2011.
13. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *J. Artif. Intell. Res.*, 2002, **16**, 321-357.
14. A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah and N. Y. Yahaya, Editon edn., 2011, vol. 188 CCIS, pp. 537-545.
15. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
16. *R version 2.9.0*, The R Foundation for Statistical Computing, 2009.
17. G. A. Biagini, N. Fisher, N. Berry, P. A. Stocks, B. Meunier, D. P. Williams, R. Bonar-Law, P. G. Bray, A. Owen, P. M. O'Neill and S. A. Ward, *Mol. Pharmacol.*, 2008, **73**, 1347-1355.
18. P. M. O'Neill, A. E. Shone, D. Stanford, G. Nixon, E. Asadollahy, B. K. Park, J. L. Maggs, P. Roberts, P. A. Stocks, G. Biagini, P. G. Bray, J. Davies, N. Berry, C. Hall, K. Rimmer, P. A. Winstanley, S. Hindley, R. B. Bambal, C. B. Davis, M. Bates, S. L. Gresham, R. A. Brigandi, F. M. Gomez-de-las-Heras, D. V. Gargallo, S. Parapini, L. Vivas, H. Lander, D. Taramelli and S. A. Ward, *Journal of Medicinal Chemistry*, 2009, **52**, 1828-1844.
19. M. Smilkstein, N. Sriwilaijaroen, J. X. Kelly, P. Wilairat and M. Riscoe, *Antimicrob. Agents Chemother.*, 2004, **48**, 1803-1806.
20. W. Trager and J. B. Jenson, *Nature*, 1978, **273**, 621-622.
21. W. Trager and J. B. Jensen, *SCIENCE*, 1976, **193**, 673-675.
22. C. Lambros and J. P. Vanderberg, *J. Parasitol.*, 1979, **65**, 418-420.
23. N. Fisher, R. Abd Majid, T. Antoine, M. Al-Helal, A. J. Warman, D. J. Johnson, A. S. Lawrenson, H. Ranson, P. M. O'Neill, S. A. Ward and G. A. Biagini, *The Journal of biological chemistry*, 2012, **287**, 9731-9741.
24. N. Fisher, A. J. Warman, S. A. Ward and G. A. Biagini, in *Methods in Enzymology, Vol 456*, ed. W. S. Allison, Elsevier Academic Press Inc, San Diego, Editon edn., 2009, vol. 456, pp. 303-320.
25. N. Fisher, C. K. Castleden, I. Bourges, G. Brasseur, G. Dujardin and B. Meunier, *J. Biol. Chem.*, 2004, **279**, 12951-12958.
26. Microsoft, *Microsoft Office Excel*, 2007.
27. *GOLD 5.0.1*, CCDC Software Limited 2005-2010, http://www.ccdc.cam.ac.uk/products/.
28. S. R. N. Solmaz and C. Hunte, *J. Biol. Chem.*, 2008, **283**, 17542-17549.
29. M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, J. W. M. Nissink, R. D. Taylor and R. Taylor, *Journal of Medicinal Chemistry*, 2005, **48**, 6504-6515.
30. L. Esser, B. Quinn, Y. F. Li, M. Q. Zhang, M. Elberry, L. Yu, C. A. Yu and D. Xia, *Journal of Molecular Biology*, 2004, **341**, 281-302.
31. H. Palsdottir, C. G. Lojero, B. L. Trumpower and C. Hunte, *J. Biol. Chem.*, 2003, **278**, 31303-31311.
32. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *Journal of Molecular Biology*, 1997, **267**, 727-748.
33. G. Jones, P. Willett and R. C. Glen, *Journal of Molecular Biology*, 1995, **245**, 43-53.

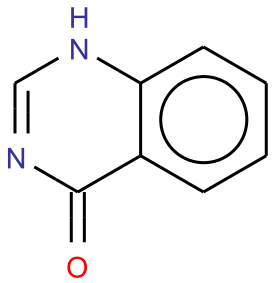
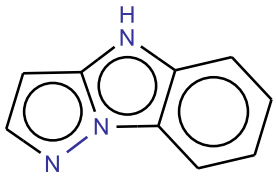
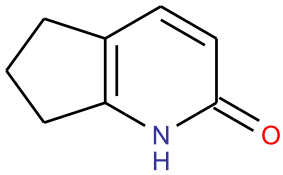
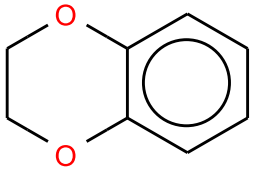
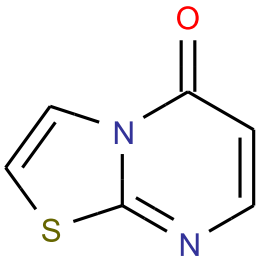
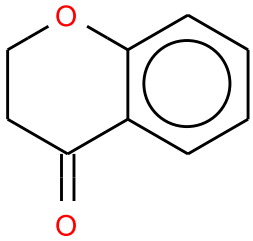
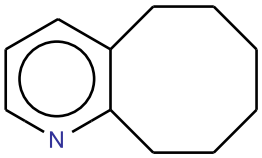
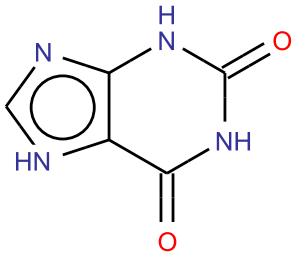
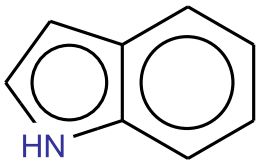
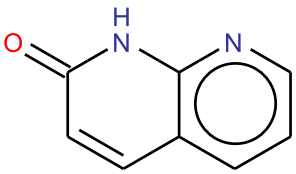
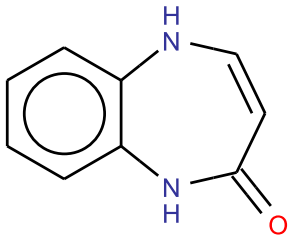
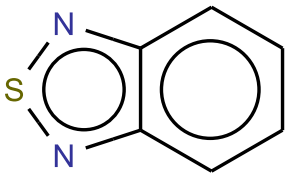
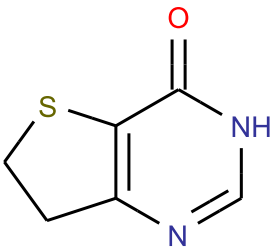
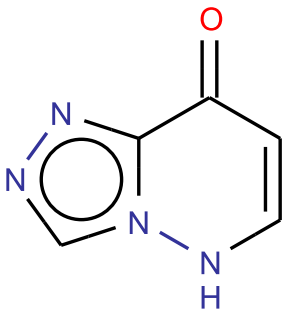
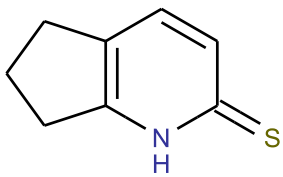
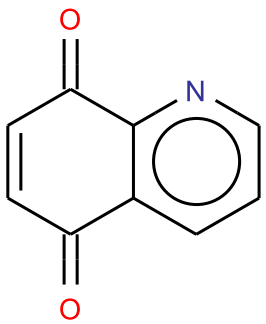
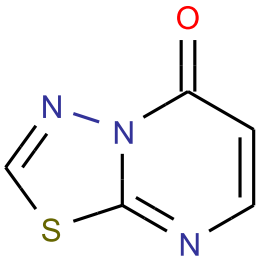
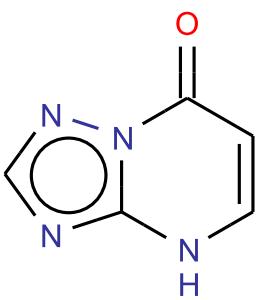
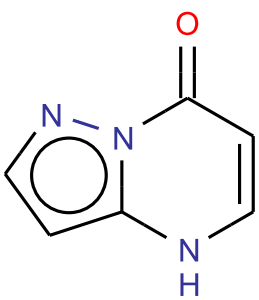
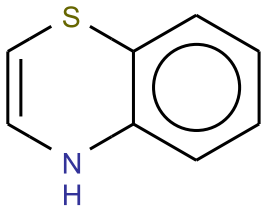
34. C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead and M. D. Eldridge, *Proteins*, 1998, **33**, 367-382.
35. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 425-445.
36. C. W. Murray, T. R. Auton and M. D. Eldridge, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 503-519.
37. Spartan, *Wavefunction, INC*, 2008.
38. T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490-519.
39. *Spartan '08 - Tutorial and User's Guide*, Wavefunction, Inc., 2006-2009.
40. Y. Chun Wei, *PHAKISO - Pharmacokinetics In Silico*, http://www.phakiso.com/.
41. D. J. Livingstone and E. Rahr, *Quant. Struct.-Act. Relat.*, 1989, **8**, 103-108.
42. OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, Paris, 2007.
43. M. A. Salamah, *Linear Regression Analysis using Excel*, http://faculty.kfupm.edu.sa/SE/salamah/mis/linear_regression_analysis_using_Excel.htm, Accessed 30/06/09.
44. A. Golbraikh and A. Tropsha, *J. Mol. Graph.*, 2002, **20**, 269-276.
45. A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Comb. Sci.*, 2003, **22**, 69-77.
46. L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikstrom and S. Wold, *Multi- and Megavariate Data Analysis: Basic Principles and Applications*, Umetrics, 2006.
47. D. C. Whitley, M. G. Ford and D. J. Livingstone, *Journal of Chemical Information and Computer Sciences*, 2000, **40**, 1160-1168.
48. *http://www.cygwin.com/*.
49. C. W. Hsu, C. C. Chang and C. J. Lin, *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, Accessed 02/02/2012.
50. T. H. Ghaddar, J. F. Wishart, J. P. Kirby, J. K. Whitesell and M. A. Fox, *Journal of the American Chemical Society*, 2001, **123**, 12832-12836.
51. M. E. Kuehne, D. M. Roland and R. Hafter, *The Journal of Organic Chemistry*, 1978, **43**, 3705-3710.
52. A. L. Pumphrey, H. Dong and T. G. Driver, *Angewandte Chemie International Edition*, 2012, **51**, 5920-5923.
53. J. Lehmann and D. Heineke, *Archiv der Pharmazie*, 1994, **327**, 715-720.
54. J.-F. Carniaux, C. Kan-Fan, J. Royer and H.-P. Husson, *Synlett*, 1999, **1999**, 563-564.

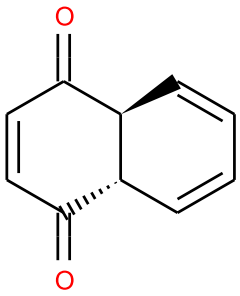
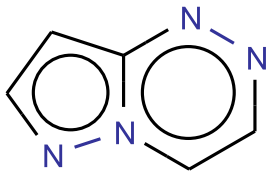
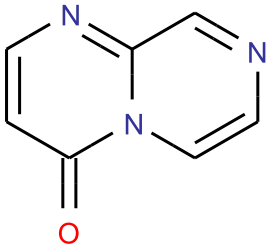
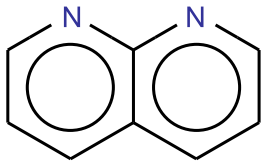
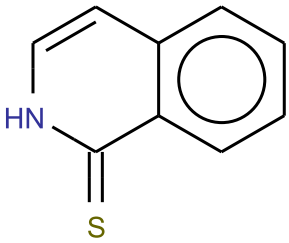
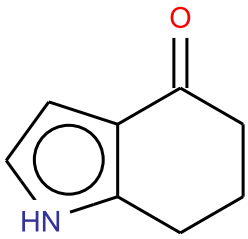
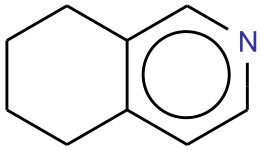
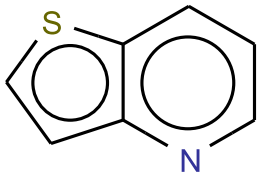
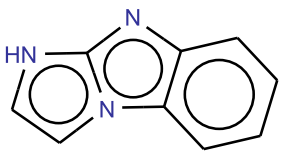
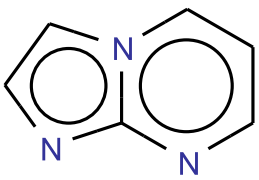
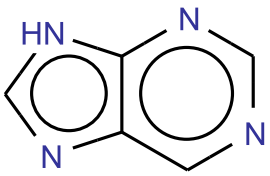
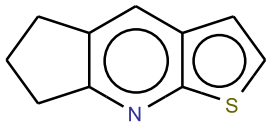
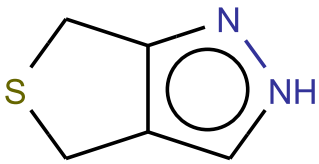
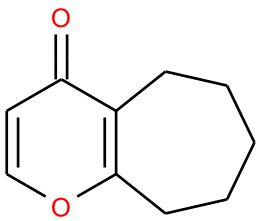
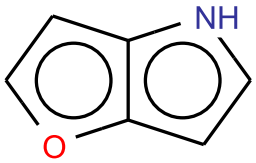
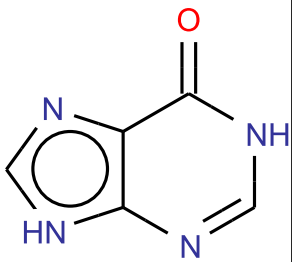
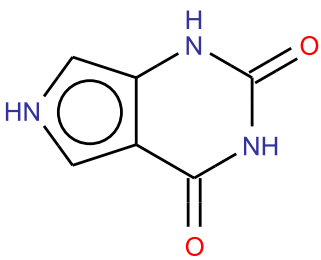
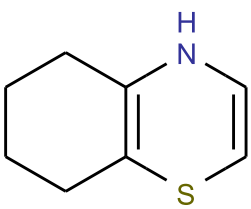
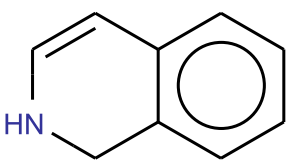
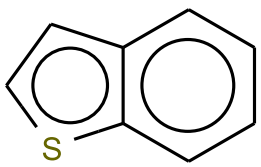
Appendix

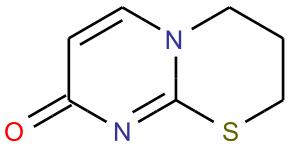
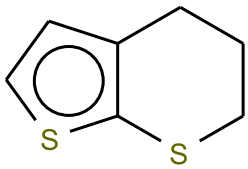
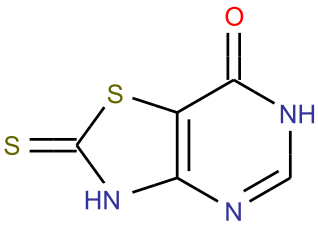
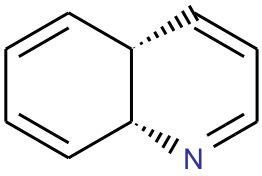
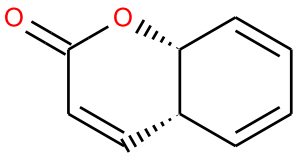
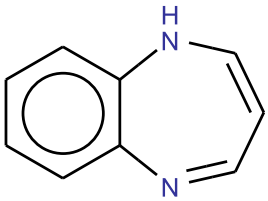
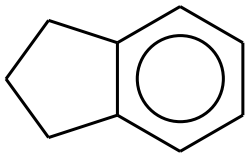
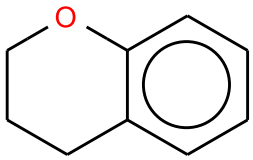
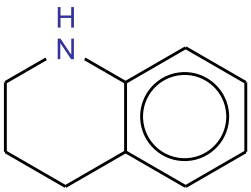
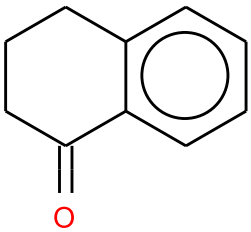
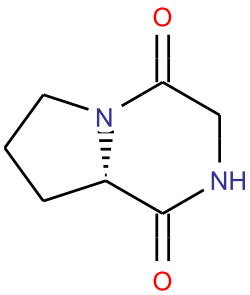
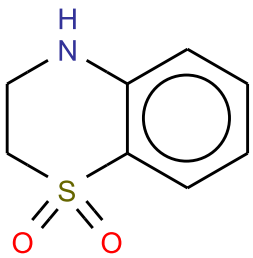
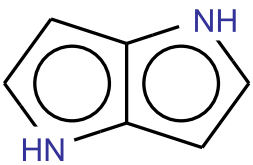
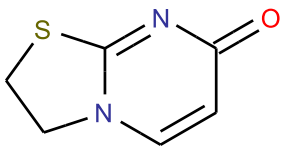
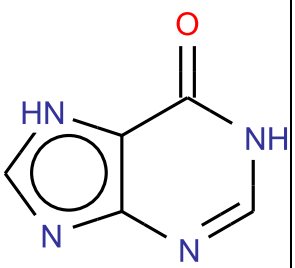
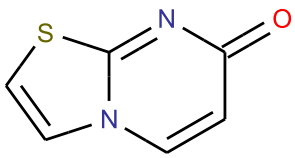
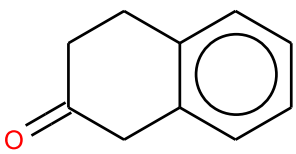
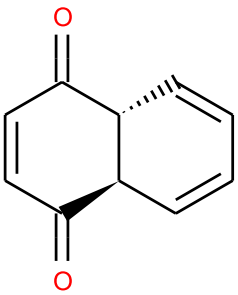
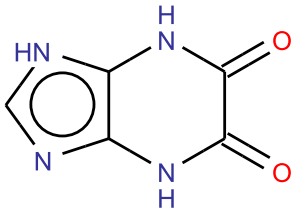
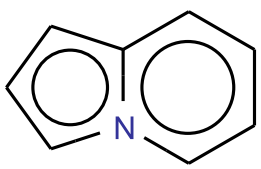
166 Bioisostere Structures used during Bioisostere Substructure Searching

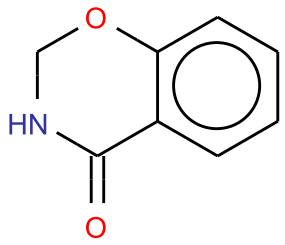
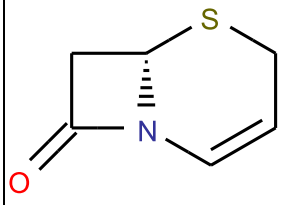
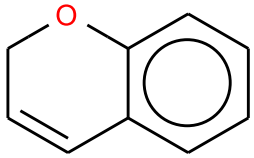
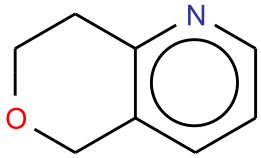
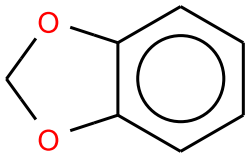
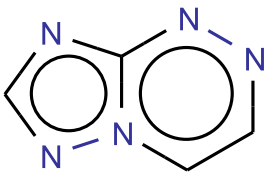
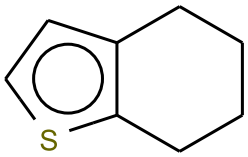
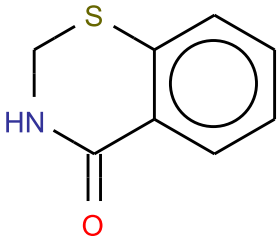
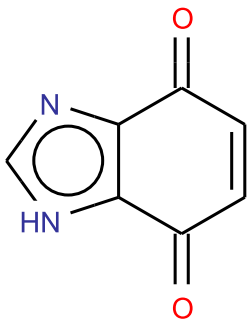
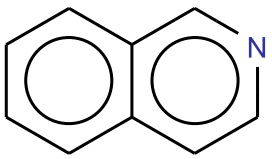
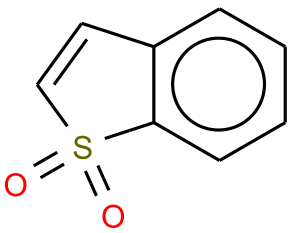
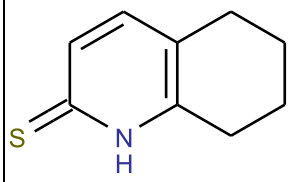
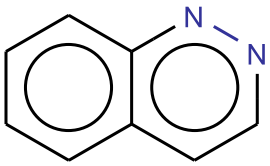
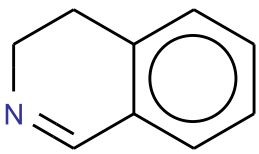
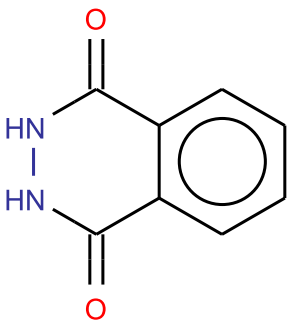
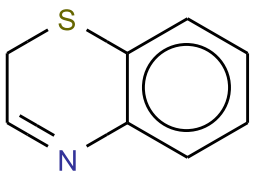
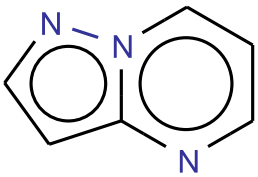
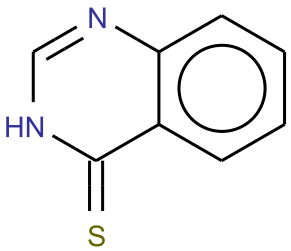
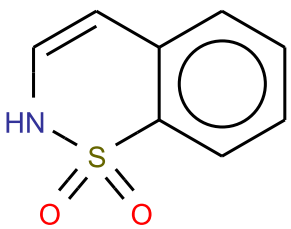
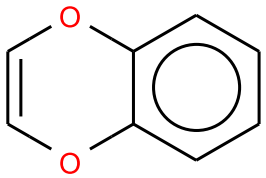
1	2	3	4
			
5	6	7	8
			
9	10	11	12
			
13	14	15	16
			
17	18	19	20
			

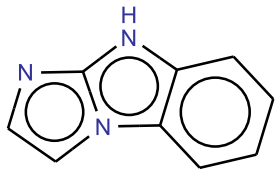
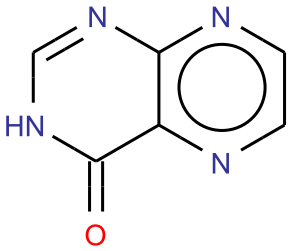
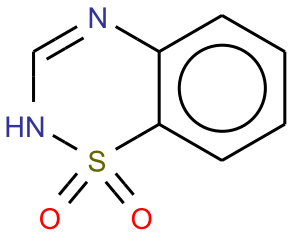
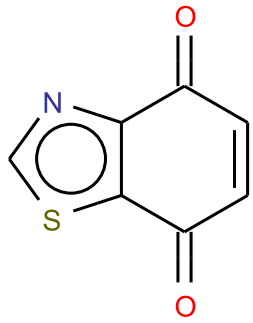
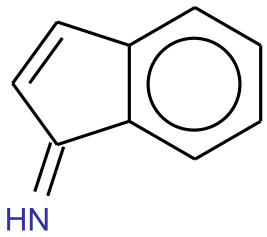
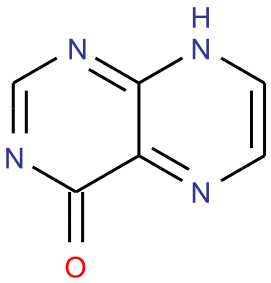
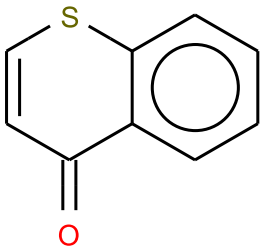
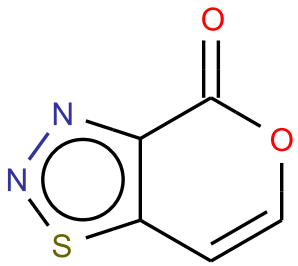
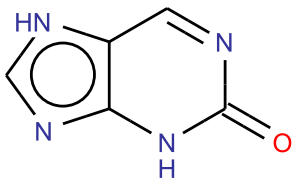
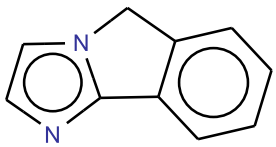
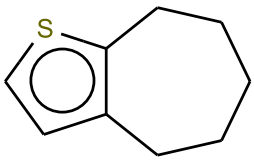
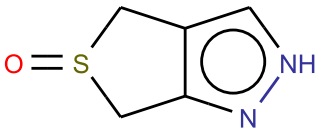
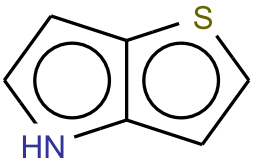
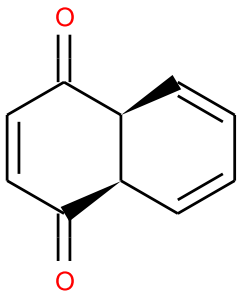
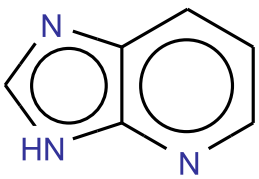
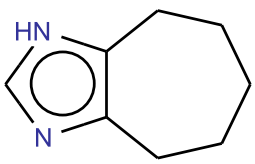
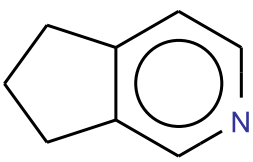
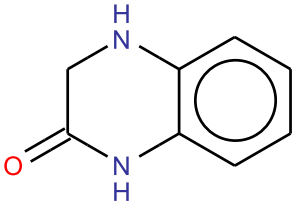
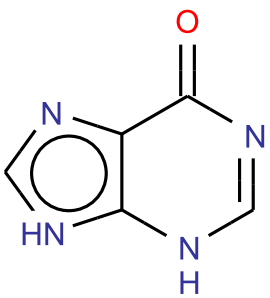
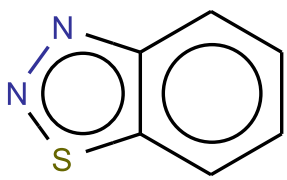
21	22	23	24
			
25	**26**	**27**	**28**
			
29	**30**	**31**	**32**
			
33	**34**	**35**	**36**
			
37	**38**	**39**	**40**
			

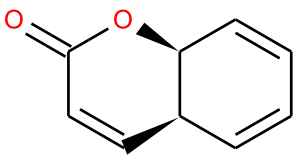
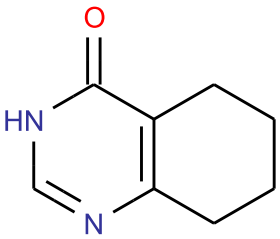
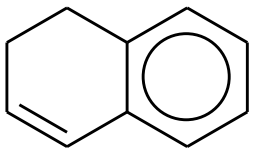
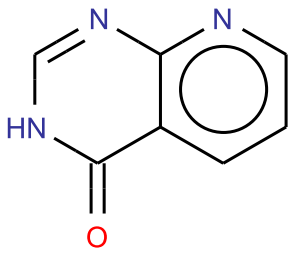
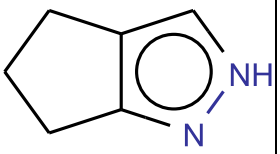
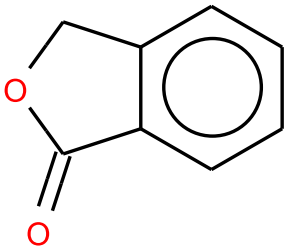
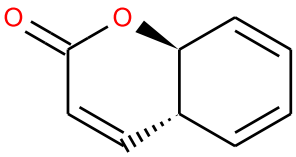
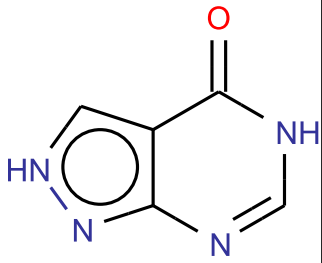
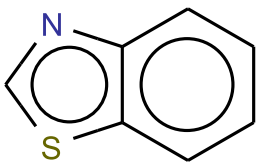
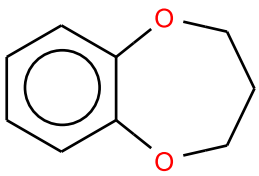
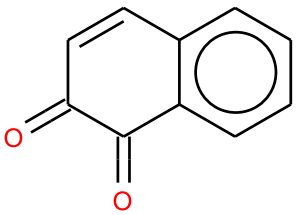
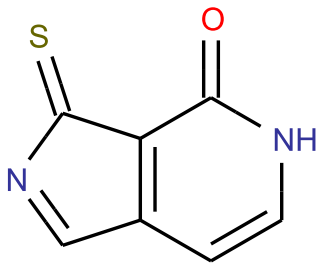
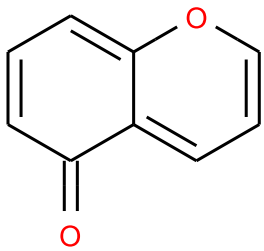
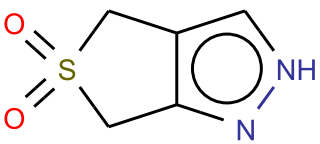
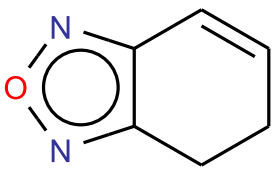
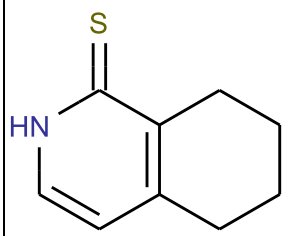
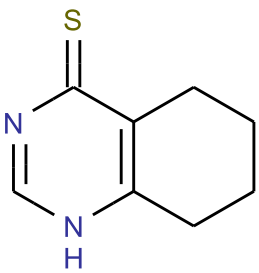
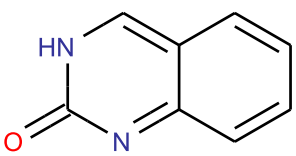
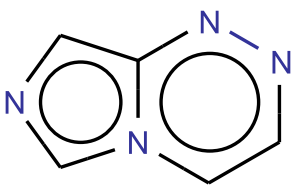
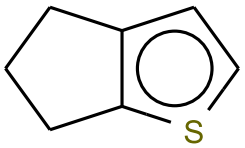
41	42	43	44
			
45	**46**	**47**	**48**
			
49	**50**	**51**	**52**
			
53	**54**	**55**	**56**
			
57	**58**	**59**	**60**
			

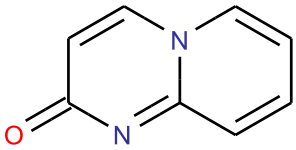
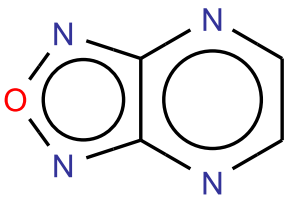
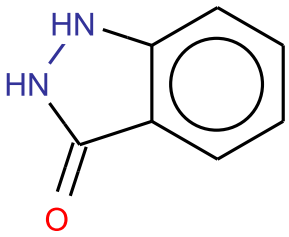
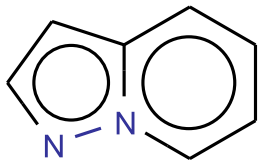
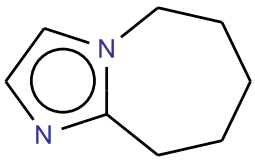
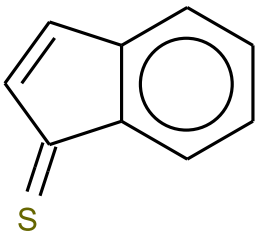
61	62	63	64
			
65	**66**	**67**	**68**
			
69	**70**	**71**	**72**
			
73	**74**	**75**	**76**
			
77	**78**	**79**	**80**
			

81	82	83	84
			
85	**86**	**87**	**88**
			
89	**90**	**91**	**92**
			
93	**94**	**95**	**96**
			
97	**98**	**99**	**100**
			

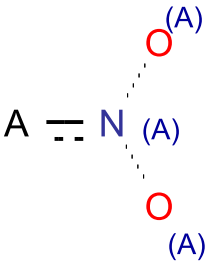
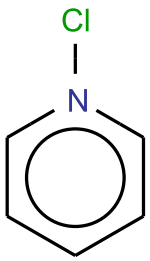
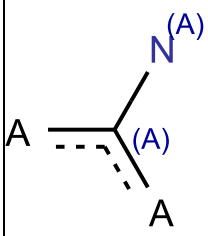
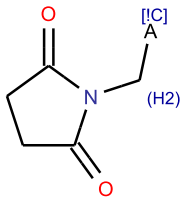
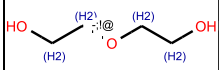
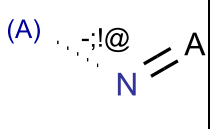
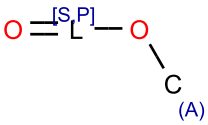
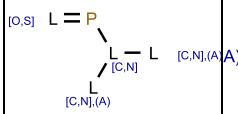
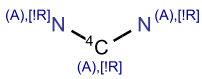
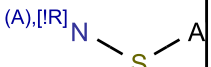
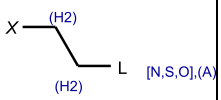
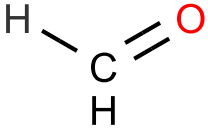
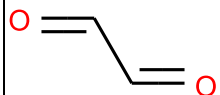
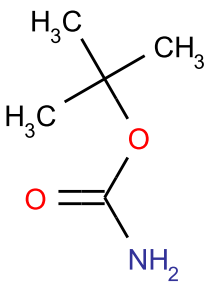
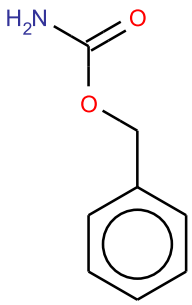
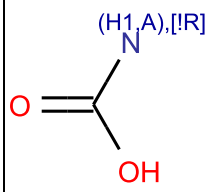
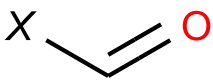
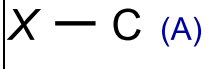
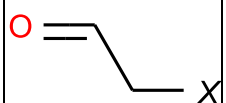
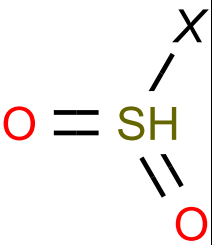
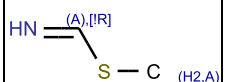
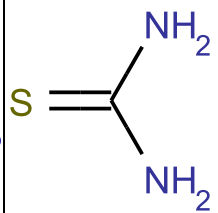
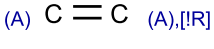
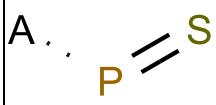
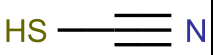
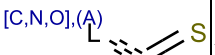
101	102	103	104
			
105	**106**	**107**	**108**
			
109	**110**	**111**	**112**
			
113	**114**	**115**	**116**
			
117	**118**	**119**	**120**
			

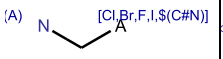
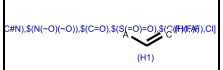
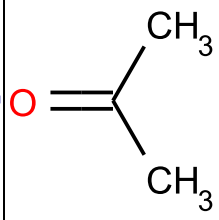
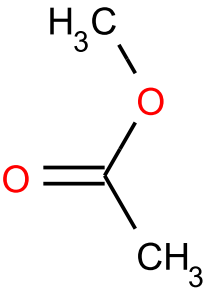
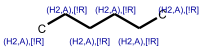
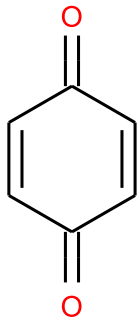
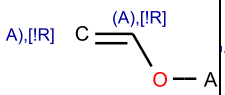
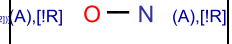
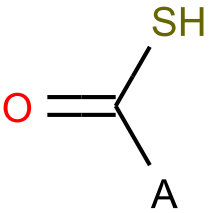
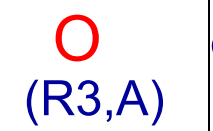
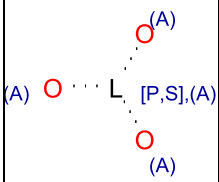
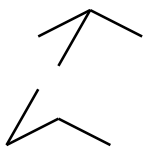
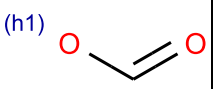
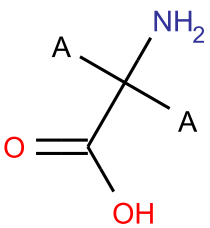
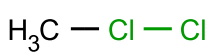
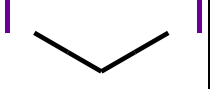
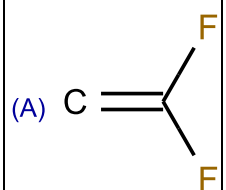
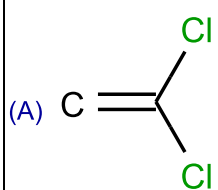
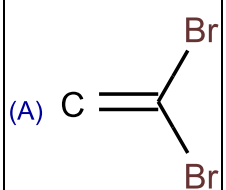
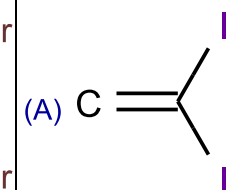
121	122	123	124
			
125	126	127	128
			
129	130	131	132
			
133	134	135	136
			
137	138	139	140
			

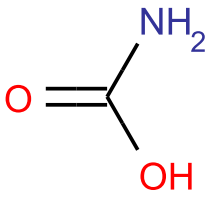
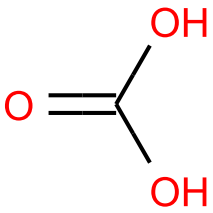
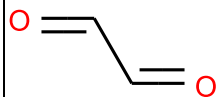
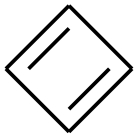
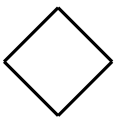
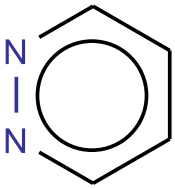
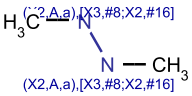
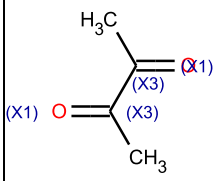
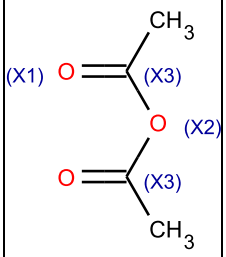
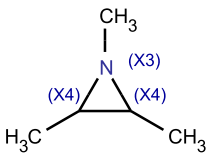
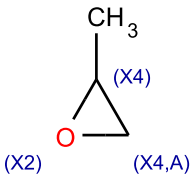
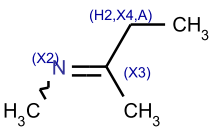
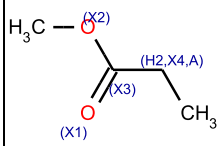
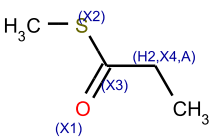
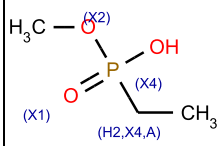
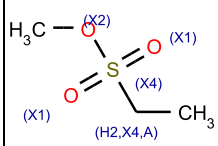
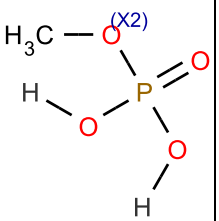
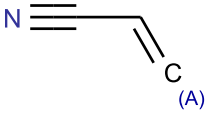
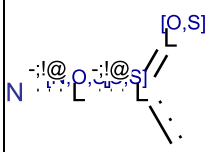
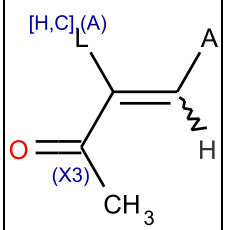
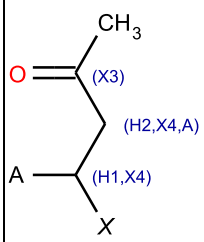
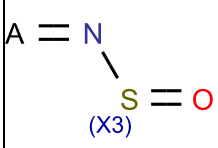
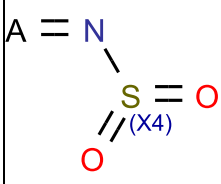
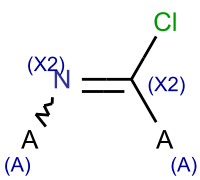
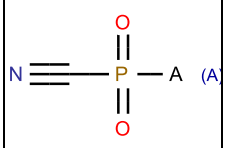
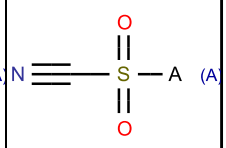
141	142	143	144
			
145	**146**	**147**	**148**
			
149	**150**	**151**	**152**
			
153	**154**	**155**	**156**
			
157	**158**	**159**	**160**
			

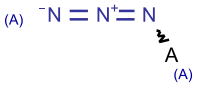
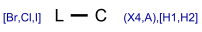
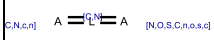
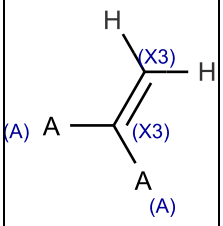
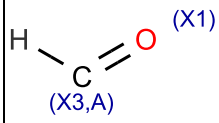
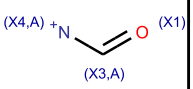
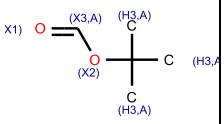
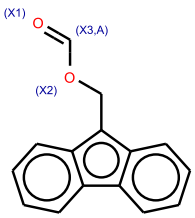
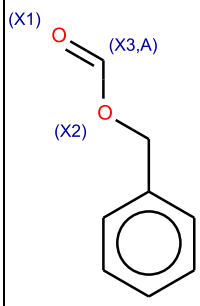
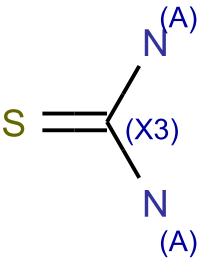
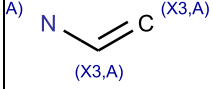
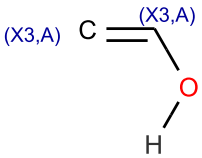
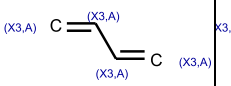
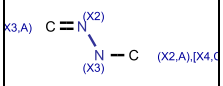
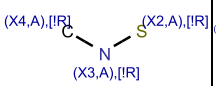
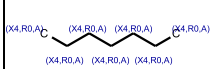
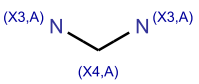
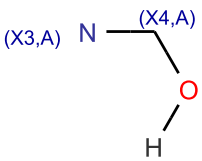
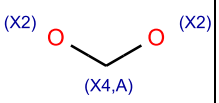
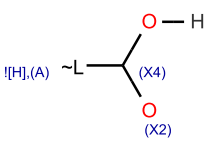
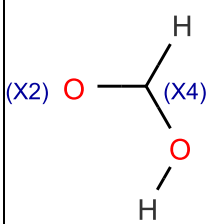
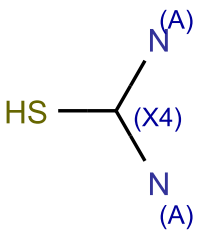
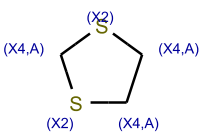
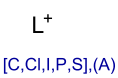
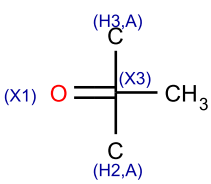
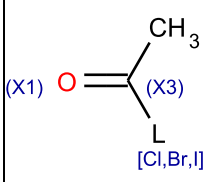
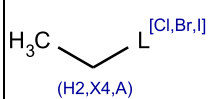
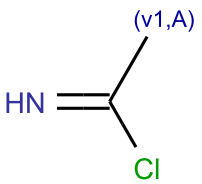
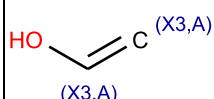
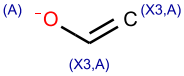
161	162	163	164
			
165	**166**		
			

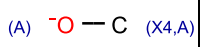
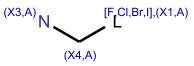
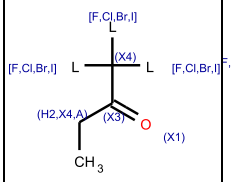
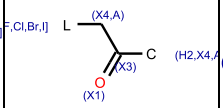
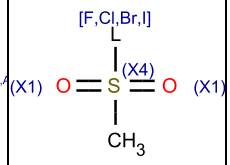
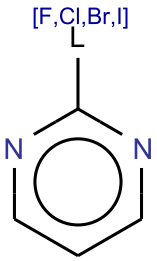
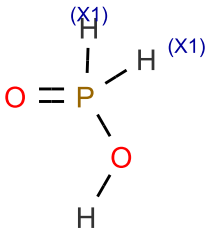
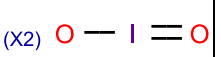
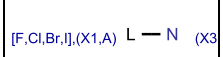
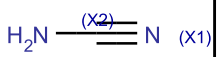
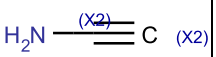
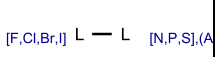
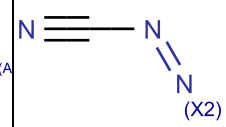
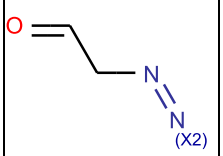
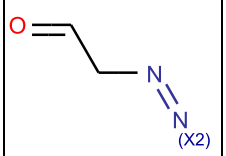
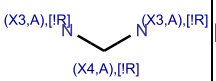
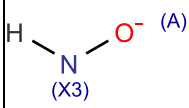
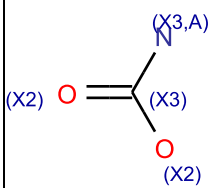
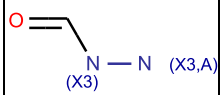
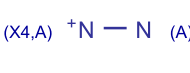
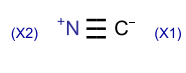
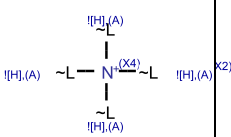
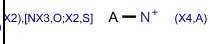
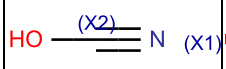
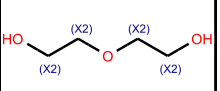
274 Toxicophores taken from the Literature and used during Compound Filtering

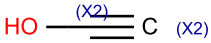
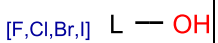
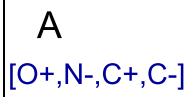
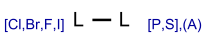
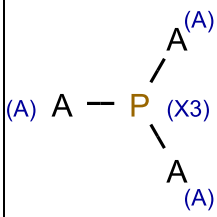
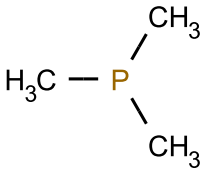
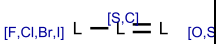
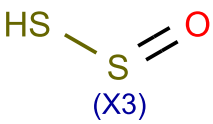
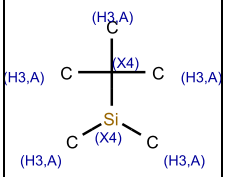
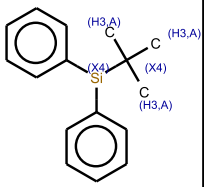
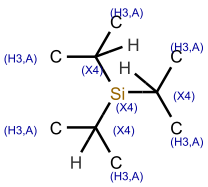
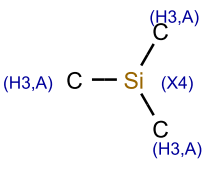
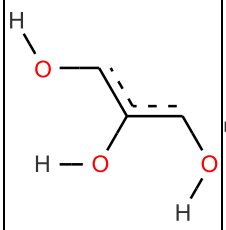
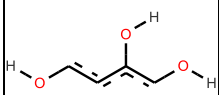
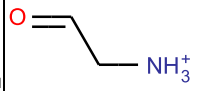
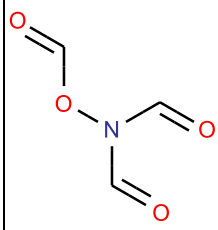
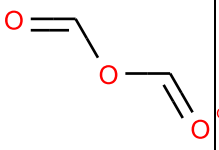
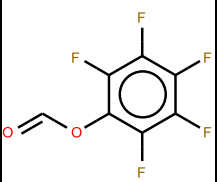
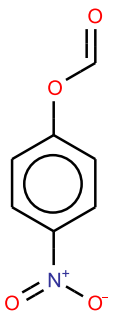
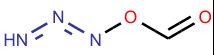
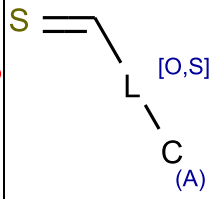
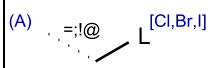
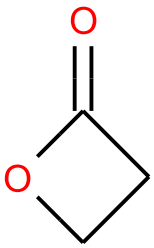
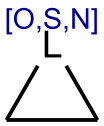
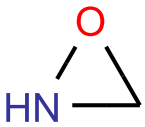
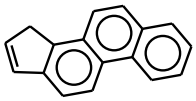
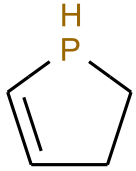
1	2	3	4	5	6
			L [Si,B,Se],(A)		
7	**8**	**9**	**10**	**11**	**12**
					
13	**14**	**15**	**16**	**17**	**18**
					
19	**20**	**21**	**22**	**23**	**24**
					
25	**26**	**27**	**28**	**29**	**30**
			S (H1,A)		

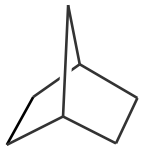
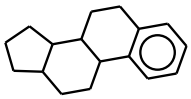
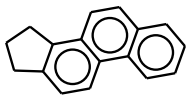
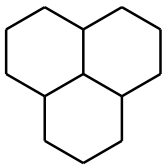
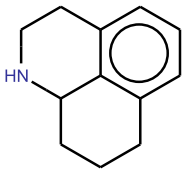
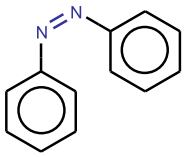
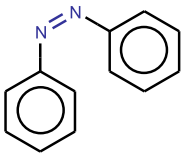
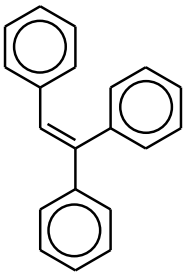
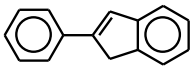
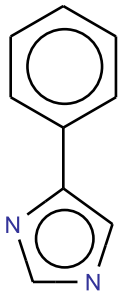
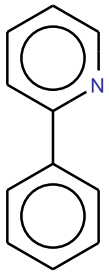
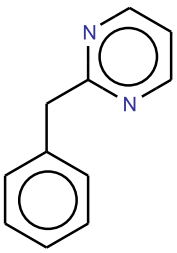
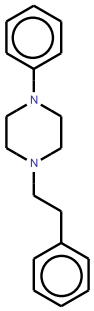
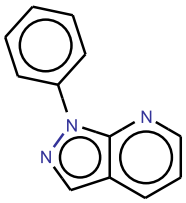
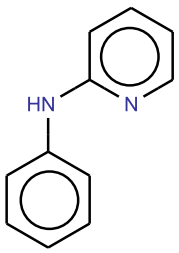
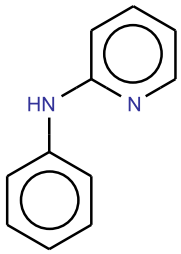
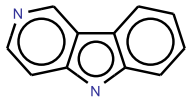
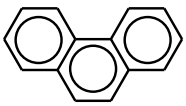
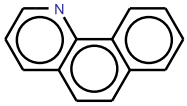
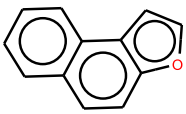
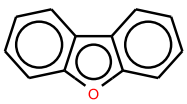
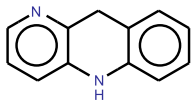
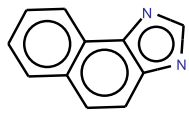
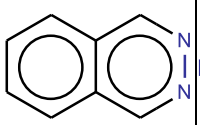
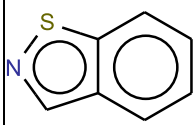
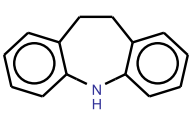
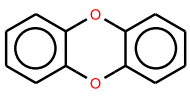
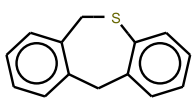
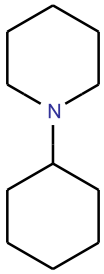
31	32	33	34	35	36
					
37	**38**	**39**	**40**	**41**	**42**
					
43	**44**	**45**	**46**	**47**	**48**
					
49	**50**	**51**	**52**	**53**	**54**
					
55	**56**	**57**	**58**	**59**	**60**
					

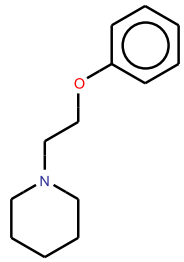
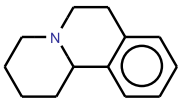
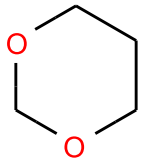
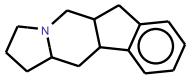
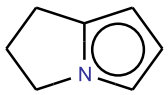
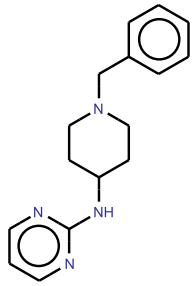
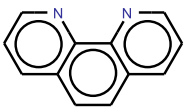
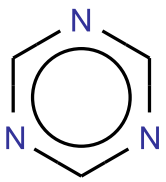
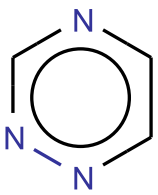
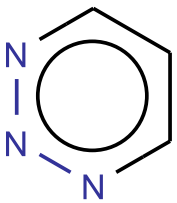
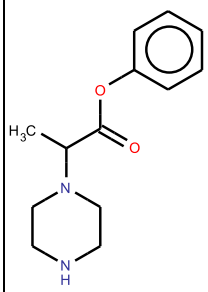
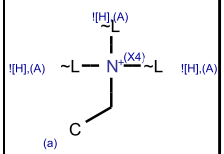
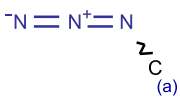
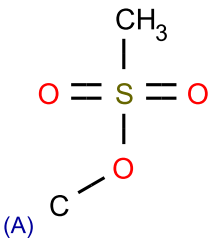
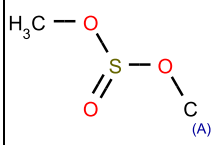
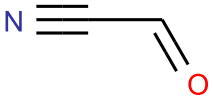
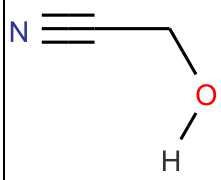
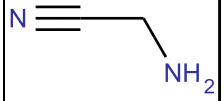
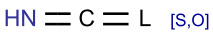
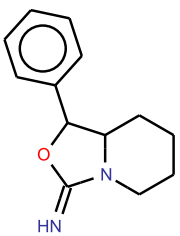
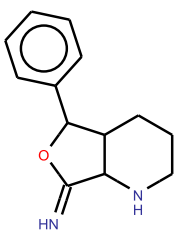
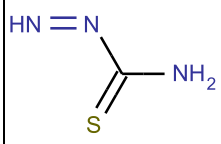
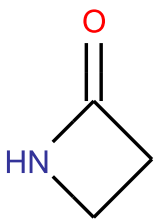
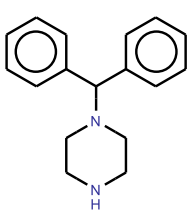
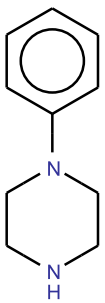
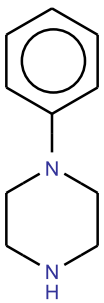
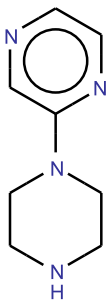
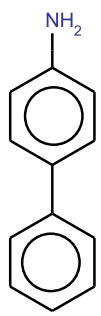
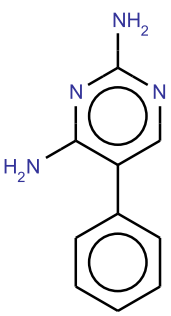
61	62	63	64	65	66
					
67	**68**	**69**	**70**	**71**	**72**
					
73	**74**	**75**	**76**	**77**	**78**
					
79	**80**	**81**	**82**	**83**	**84**
					
85	**86**	**87**	**88**	**89**	**90**
					

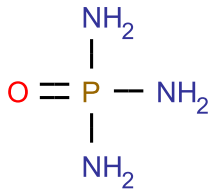
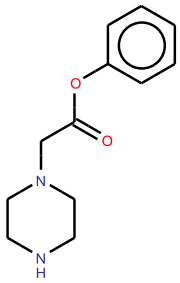
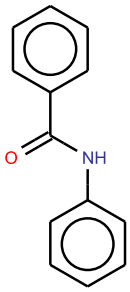
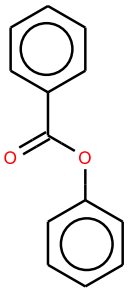
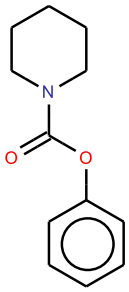
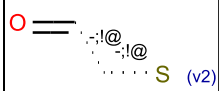
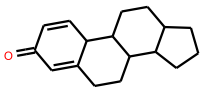
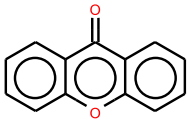
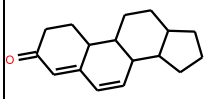
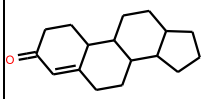
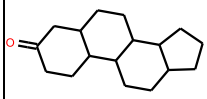
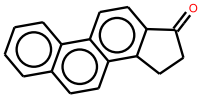
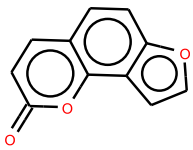
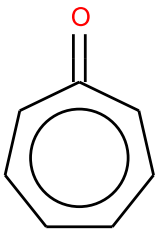
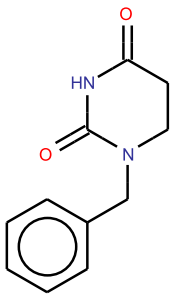
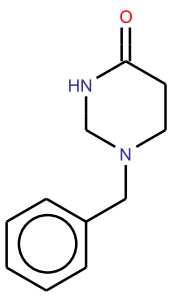
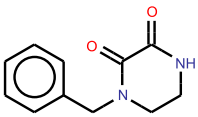
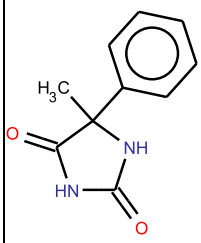
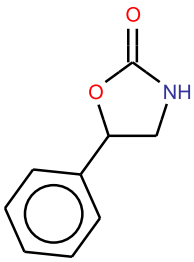
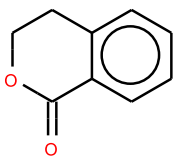
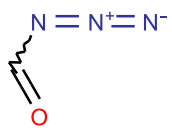
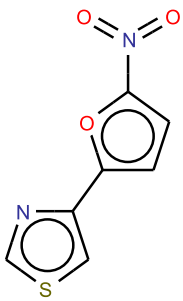
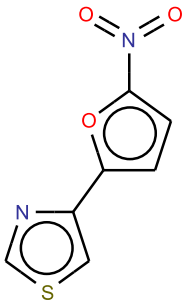
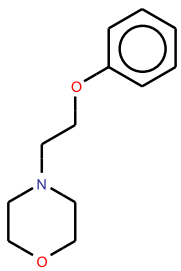
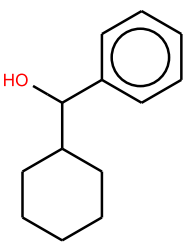
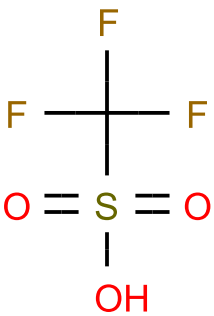
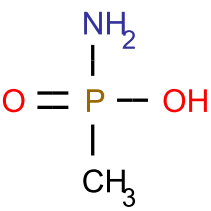
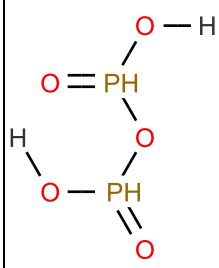
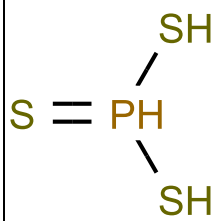
91	92	93	94	95	96
					
97	**98**	**99**	**100**	**101**	**102**
					
103	**104**	**105**	**106**	**107**	**108**
					
109	**110**	**111**	**112**	**113**	**114**
					
115	**116**	**117**	**118**	**119**	**120**
					

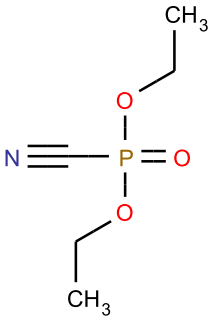
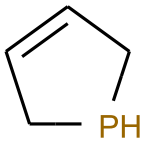
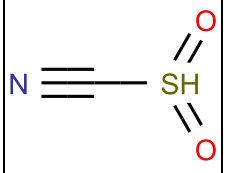
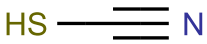
121	122	123	124	125	126
					
127	**128**	**129**	**130**	**131**	**132**
					
133	**134**	**135**	**136**	**137**	**138**
					
139	**140**	**141**	**142**	**143**	**144**
					
145	**146**	**147**	**148**	**149**	**150**
					

151	152	153	154	155	156
					
157	**158**	**159**	**160**	**161**	**162**
					
163	**164**	**165**	**166**	**167**	**168**
					
169	**170**	**171**	**172**	**173**	**174**
					
175	**176**	**177**	**178**	**179**	**180**
					

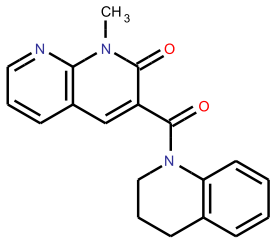
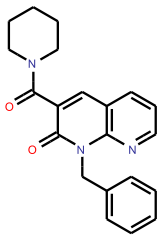
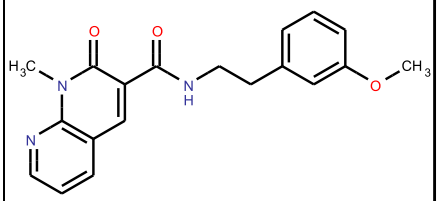
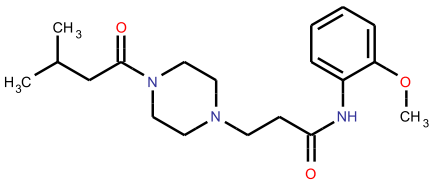
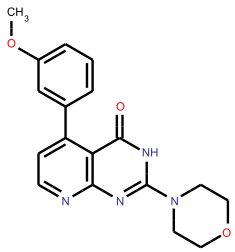
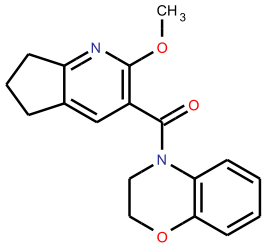
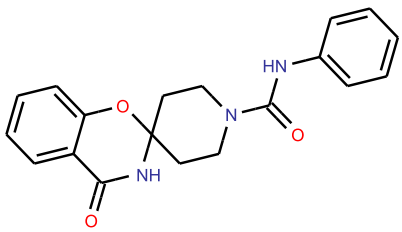
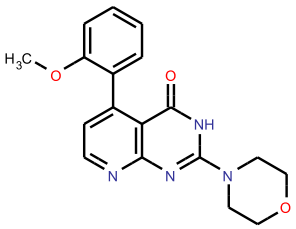
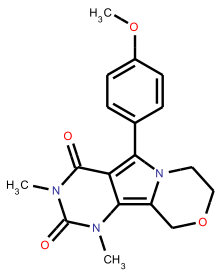
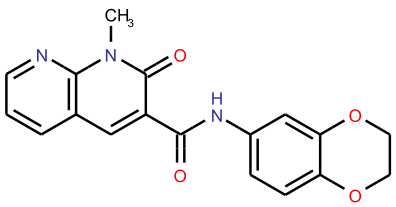
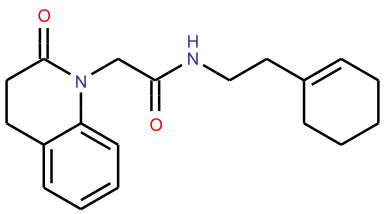
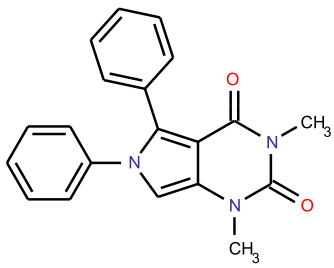
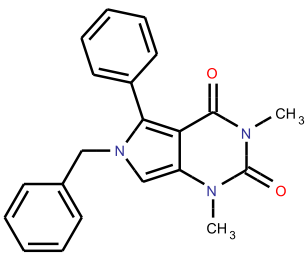
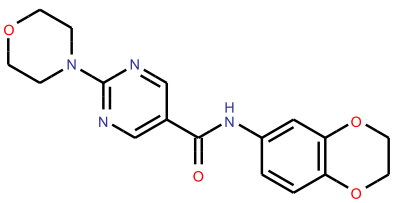
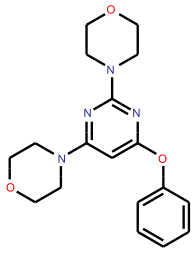
181	182	183	184	185	186
					
187	**188**	**189**	**190**	**191**	**192**
					
193	**194**	**195**	**196**	**197**	**198**
					
199	**200**	**201**	**202**	**203**	**204**
					
205	**206**	**207**	**208**	**209**	**210**
					

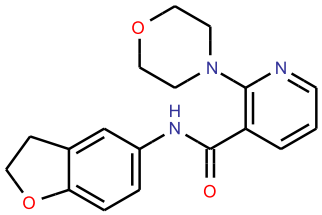
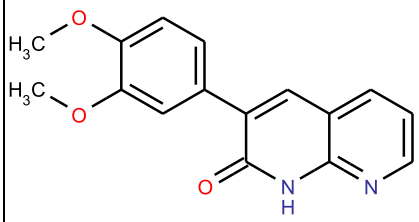
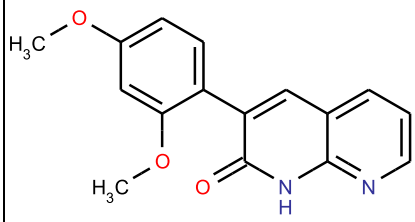
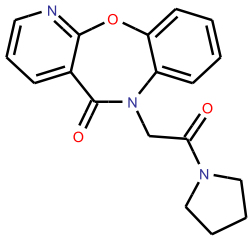
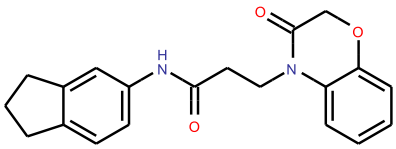
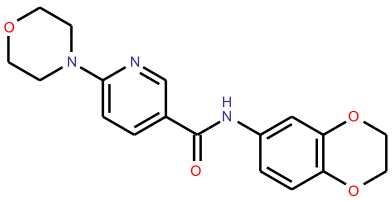
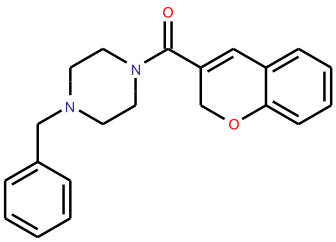
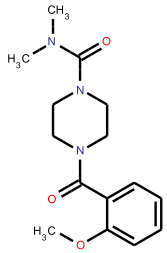
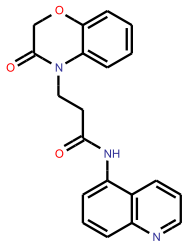
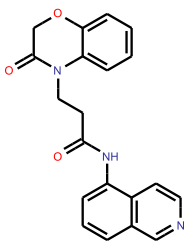
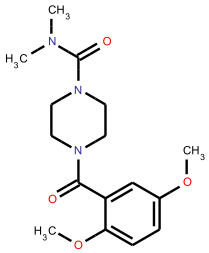
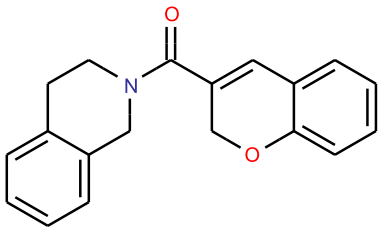
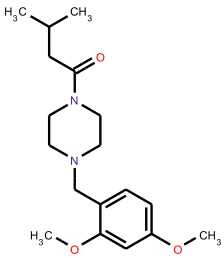
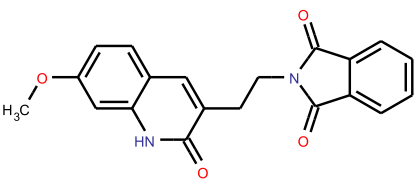
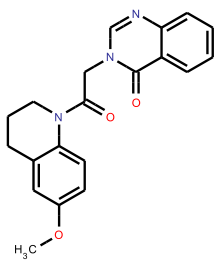
211	212	213	214	215	216
					
217	**218**	**219**	**220**	**221**	**222**
					
223	**224**	**225**	**226**	**227**	**228**
					
229	**230**	**231**	**232**	**233**	**234**
					
235	**236**	**237**	**238**	**239**	**240**
					

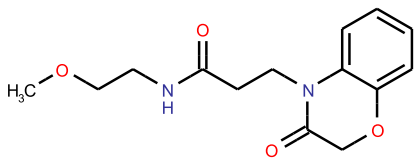
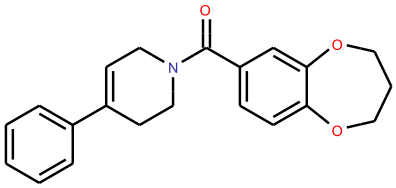
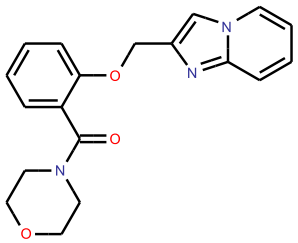
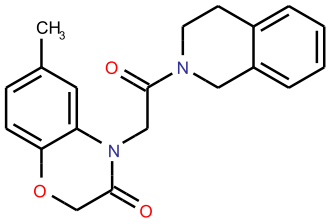
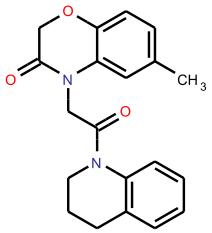
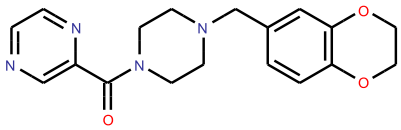
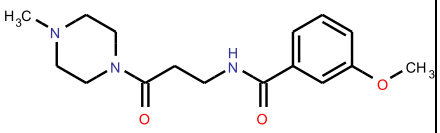
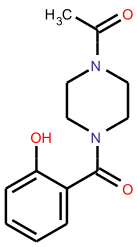
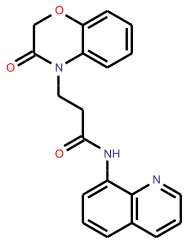
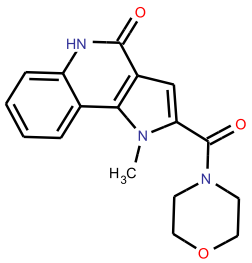
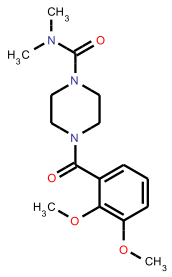
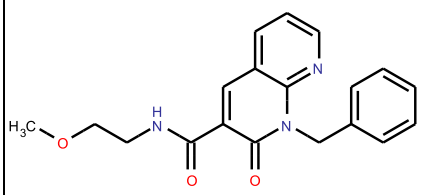
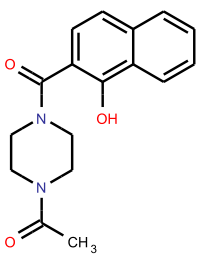
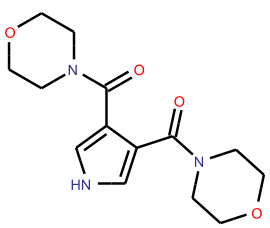
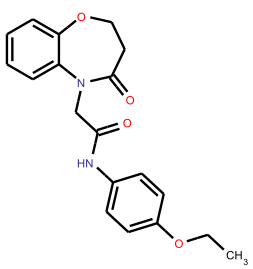
241	242	243	244	245	246
					
247	**248**	**249**	**250**	**251**	**252**
					
253	**254**	**255**	**256**	**257**	**258**
					
259	**260**	**261**	**262**	**263**	**264**
					
265	**266**	**267**	**268**	**269**	**270**
					

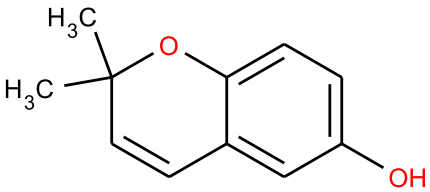
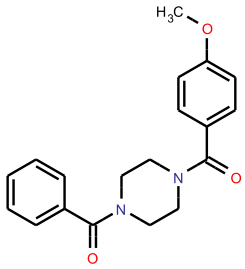
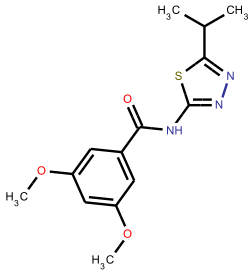
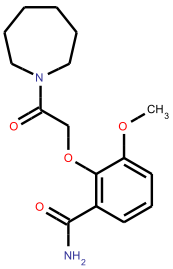
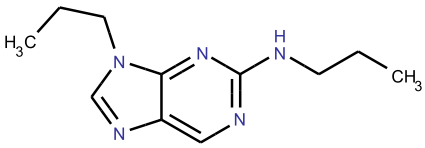
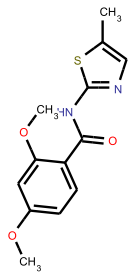
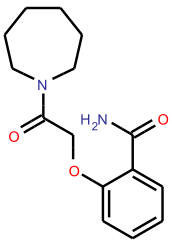
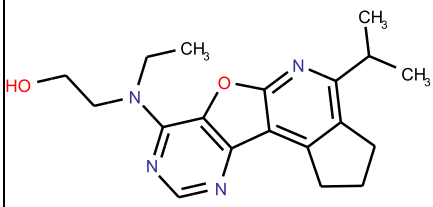
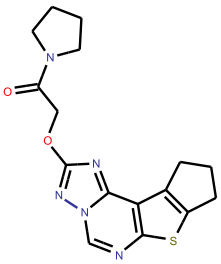
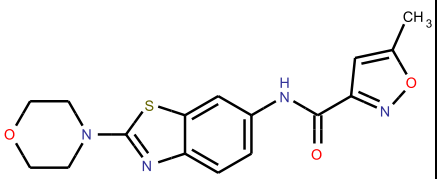
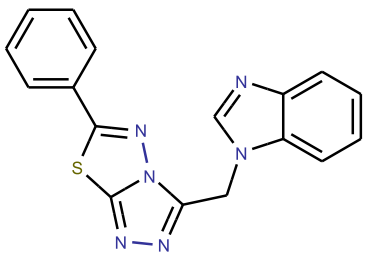
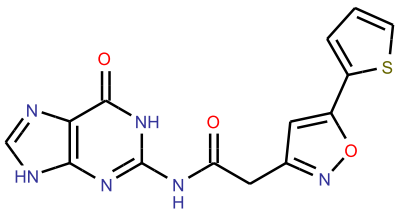
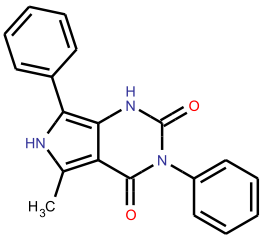
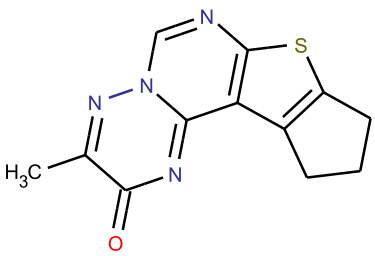
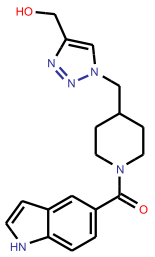
271	272	273	274		
					

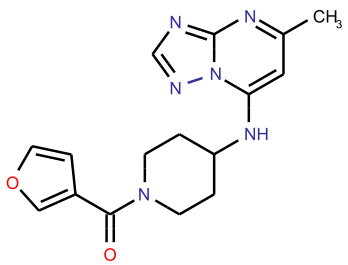
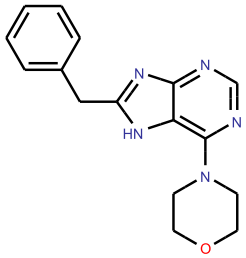
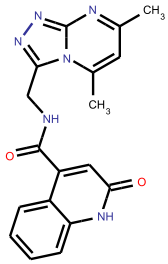
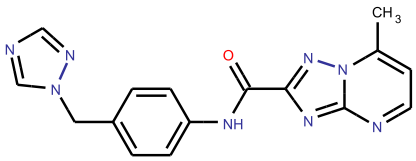
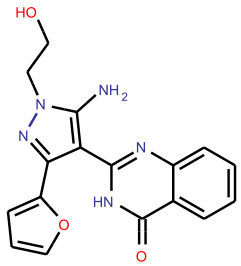
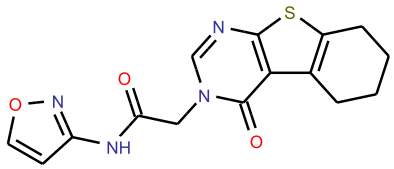
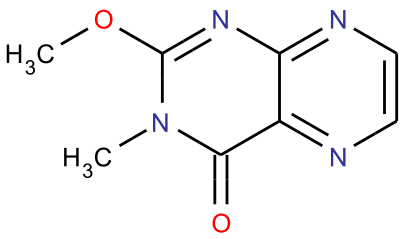
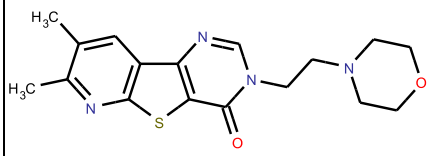
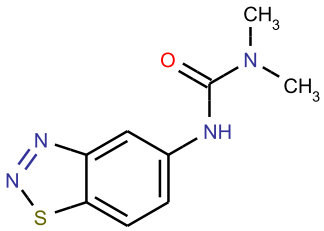
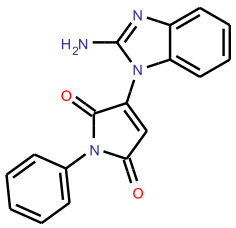
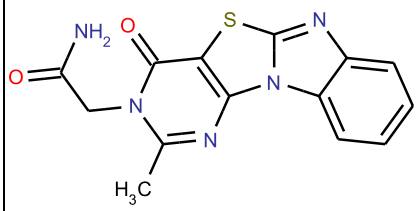
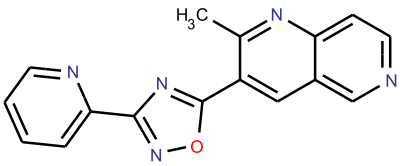
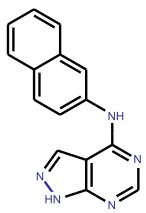
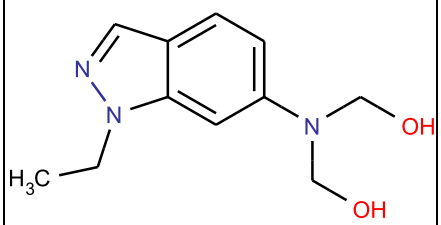
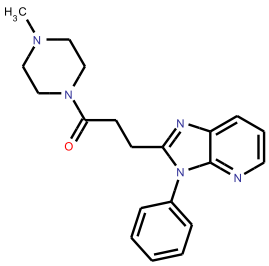
Final Selection of 139 Compounds

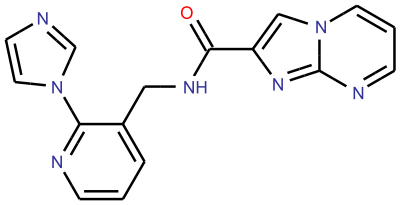
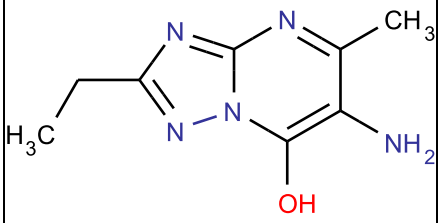
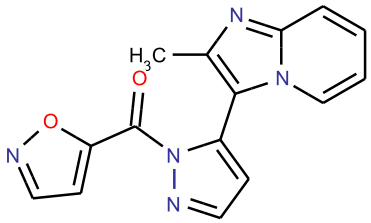
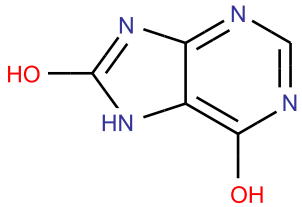
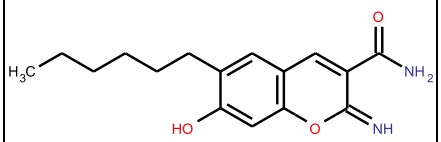
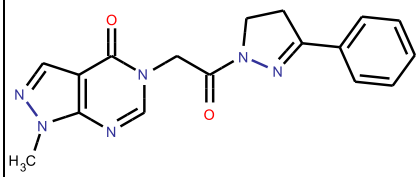
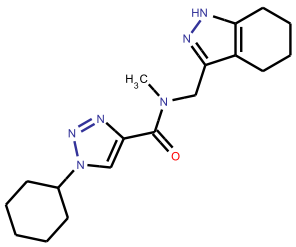
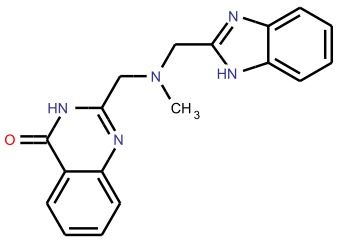
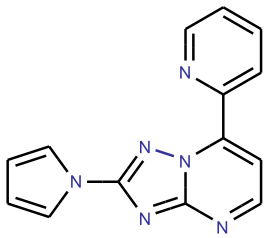
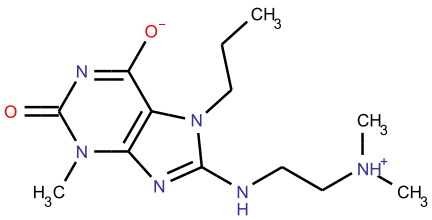
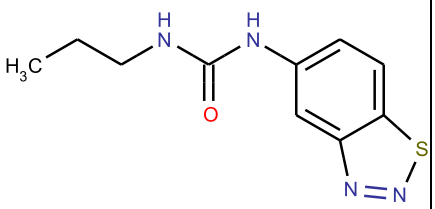
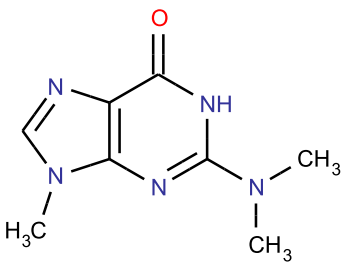
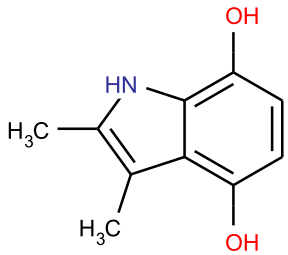
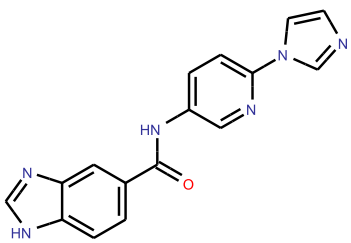
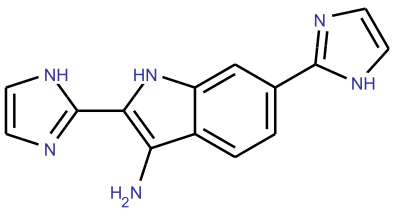
1	2	3
		
Data: ZINC05303593 Final_Score: 5.3831135902	Data: ZINC05304204 Final_Score: 5.3831135902	Data: ZINC05303747 Final_Score: 5.2611111111
4	**5**	**6**
		
Data: ZINC25530787 Final_Score: 5.1215806202	Data: ZINC12390004 Final_Score: 5.0380434782	Data: ZINC19732331 Final_Score: 5.0073529411
7	**8**	**9**
		
Data: ZINC20864976 Final_Score: 4.9930555555	Data: ZINC12391040 Final_Score: 4.9802631578	Data: ZINC04776555 Final_Score: 4.9802631578
10	**11**	**12**
		
Data: ZINC05303683 Final_Score: 4.9771126760	Data: ZINC06874526 Final_Score: 4.9499095968	Data: ZINC00206360 Final_Score: 4.9409842673
13	**14**	**15**
		
Data: ZINC00033678 Final_Score: 4.9309177806	Data: ZINC15777895 Final_Score: 4.9111607142	Data: ZINC25492425 Final_Score: 4.9104838709

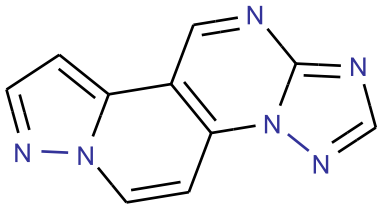
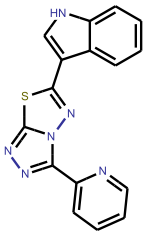
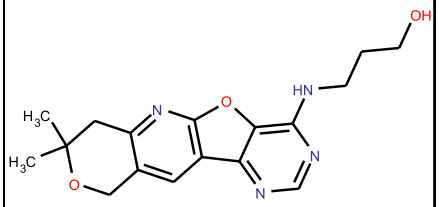
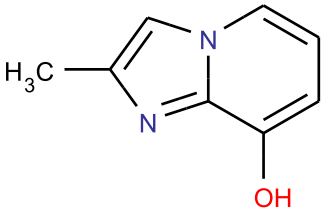
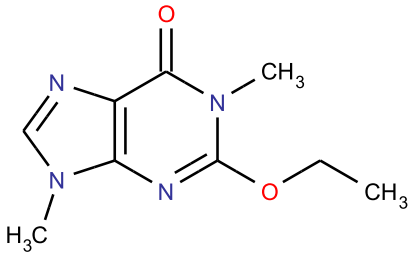
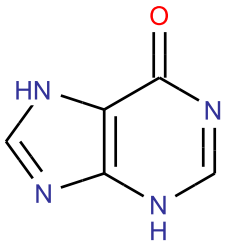
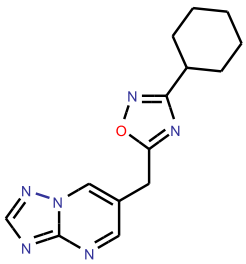
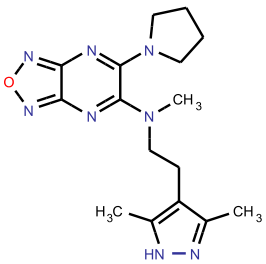
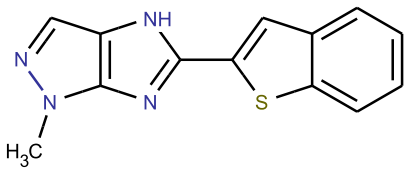
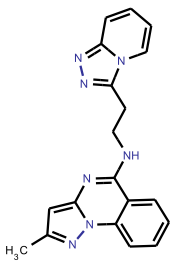
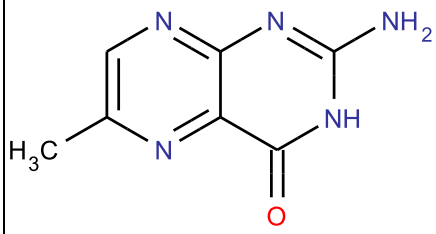
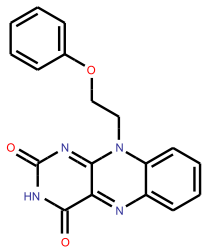
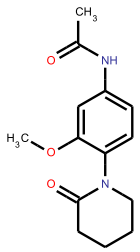
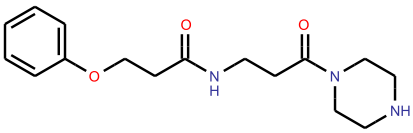
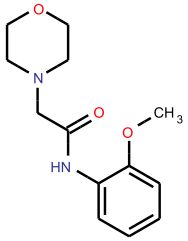
16	17	18
		
Data: ZINC27127427 Final_Score: 4.9022435897	Data: ZINC19732046 Final_Score: 4.9021739130	Data: ZINC19732050 Final_Score: 4.9021739130
19	**20**	**21**
		
Data: ZINC15679958 Final_Score: 4.8988894907	Data: ZINC08061498 Final_Score: 4.8968533422	Data: ZINC24601544 Final_Score: 4.8956349206
22	**23**	**24**
		
Data: ZINC15629486 Final_Score: 4.8948863636	Data: ZINC26937015 Final_Score: 4.8934331797	Data: ZINC12849197 Final_Score: 4.8934331797
25	**26**	**27**
		
Data: ZINC21218463 Final_Score: 4.8934331797	Data: ZINC26932562 Final_Score: 4.8934331797	Data: ZINC15627439 Final_Score: 4.8932291666
28	**29**	**30**
		
Data: ZINC19782464 Final_Score: 4.8930327868	Data: ZINC08683765 Final_Score: 4.8898110887	Data: ZINC10946644 Final_Score: 4.8889168343

31	32	33
		
Data: ZINC04687080 Final_Score: 4.8888566912	Data: ZINC09450085 Final_Score: 4.8880269704	Data: ZINC27474496 Final_Score: 4.8876262626
34	**35**	**36**
		
Data: ZINC04375208 Final_Score: 4.8869334619	Data: ZINC01362062 Final_Score: 4.8869334619	Data: ZINC23786796 Final_Score: 4.8862745098
37	**38**	**39**
		
Data: ZINC12866036 Final_Score: 4.8837822014	Data: ZINC19839588 Final_Score: 4.8832089552	Data: ZINC08942475 Final_Score: 4.8818745808
40	**41**	**42**
		
Data: ZINC06771841 Final_Score: 4.8818745808	Data: ZINC26936754 Final_Score: 4.8818745808	Data: ZINC05304264 Final_Score: 4.8581081081
43	**44**	**45**
		
Data: ZINC20260323 Final_Score: 4.7731343283	Data: ZINC08376329 Final_Score: 4.7630269704	Data: ZINC07918355 Final_Score: 4.7568745808

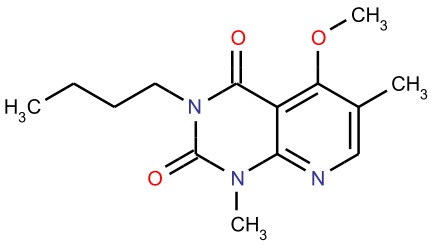
46	**47**	**48**
		
Data: ZINC01572968 Final_Score: 4.7567086161	Data: ZINC02786579 Final_Score: 4.7556318681	Data: ZINC09331449 Final_Score: 4.7384904783
49	**50**	**51**
		
Data: ZINC08394887 Final_Score: 4.6897015887	Data: ZINC01677597 Final_Score: 4.6847927728	Data: ZINC09578769 Final_Score: 4.6614406204
52	**53**	**54**
		
Data: ZINC04550317 Final_Score: 4.6412571889	Data: ZINC13206954 Final_Score: 4.6285685663	Data: ZINC12388953 Final_Score: 4.625
55	**56**	**57**
		
Data: ZINC10032083 Final_Score: 4.625	Data: ZINC00383153 Final_Score: 4.625	Data: ZINC12627737 Final_Score: 4.625
58	**59**	**60**
		
Data: ZINC01859191 Final_Score: 4.625	Data: ZINC01300129 Final_Score: 4.625	Data: ZINC23543463 Final_Score: 4.625

61  Data: ZINC24207030 Final_Score: 4.625	**62**  Data: ZINC05626381 Final_Score: 4.625	**63**  Data: ZINC25106612 Final_Score: 4.625
64  Data: ZINC24633497 Final_Score: 4.625	**65**  Data: ZINC19721738 Final_Score: 4.625	**66**  Data: ZINC14731871 Final_Score: 4.625
67  Data: ZINC01638978 Final_Score: 4.625	**68**  Data: ZINC20676655 Final_Score: 4.625	**69**  Data: ZINC00270125 Final_Score: 4.625
70  Data: ZINC20636500 Final_Score: 4.625	**71**  Data: ZINC05728724 Final_Score: 4.625	**72**  Data: ZINC19744869 Final_Score: 4.625
73  Data: ZINC04844400 Final_Score: 4.625 4.625	**74**  Data: ZINC01726887 Final_Score: 4.625	**75**  Data: ZINC21349143 Final_Score: 4.625

76	77	78
		
Data: ZINC15069834 Final_Score: 4.625	Data: ZINC12413607 Final_Score: 4.625	Data: ZINC00123941 Final_Score: 4.625
79	**80**	**81**
		
Data: ZINC13543038 Final_Score: 4.625 4.625	Data: ZINC02030238 Final_Score: 4.625	Data: ZINC10769672 Final_Score: 4.625
82	**83**	**84**
		
Data: ZINC11936087 Final_Score: 4.625	Data: ZINC23762388 Final_Score: 4.625	Data: ZINC01382940 Final_Score: 4.625
85	**86**	**87**
		
Data: ZINC19913190 Final_Score: 4.625	Data: ZINC02499601 Final_Score: 4.625	Data: ZINC08462411 Final_Score: 4.625
88	**89**	**90**
		
Data: ZINC01686243 Final_Score: 4.625	Data: ZINC24901612 Final_Score: 4.625	Data: ZINC13211971 Final_Score: 4.625

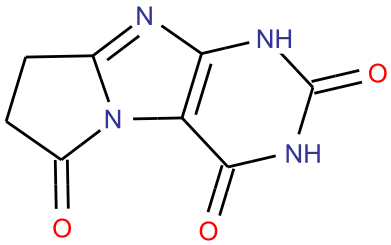
91	92	93
		
Data: ZINC12388833 Final_Score: 4.625	Data: ZINC12389909 Final_Score: 4.625	Data: ZINC02433033 Final_Score: 4.625
94	**95**	**96**
		
Data: ZINC00337579 Final_Score: 4.625	Data: ZINC00485870 Final_Score: 4.625	Data: ZINC18153302 Final_Score: 4.625
97	**98**	**99**
		
Data: ZINC20784715 Final_Score: 4.625	Data: ZINC11957975 Final_Score: 4.625	Data: ZINC20915823 Final_Score: 4.625
100	**101**	**102**
		
Data: ZINC26761855 Final_Score: 4.625	Data: ZINC00403052 Final_Score: 4.573	Data: ZINC08783555 Final_Score: 4.5679021719
103	**104**	**105**
		
Data: ZINC05441524 Final_Score: 4.5402542372	Data: ZINC20184234 Final_Score: 4.5348360655	Data: ZINC19805548 Final_Score: 4.5348360655

106



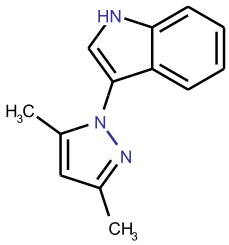
Data: ZINC09648861
Final_Score: 4.5265151515

107



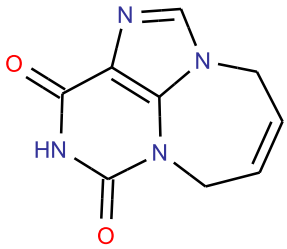
Data: ZINC26479936
Final_Score: 4.5

108



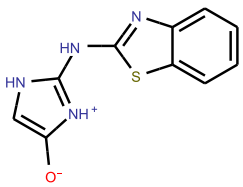
Data: ZINC04017558
Final_Score: 4.5

109



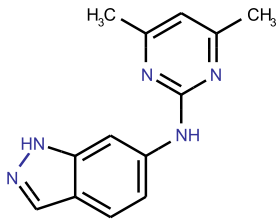
Data: ZINC01630153
Final_Score: 4.5

110



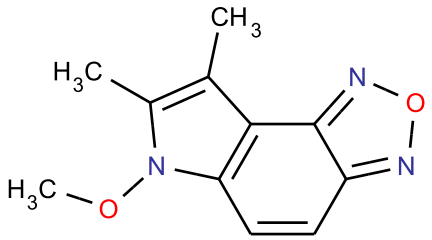
Data: ZINC08436486
Final_Score: 4.5
4.5

111



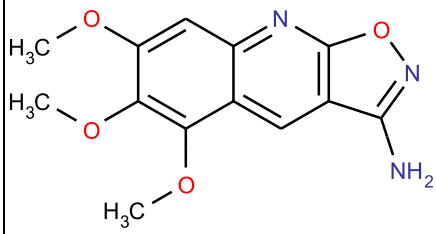
Data: ZINC08020286
Final_Score: 4.5

112



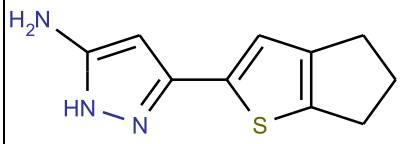
Data: ZINC02296952
Final_Score: 4.5

113



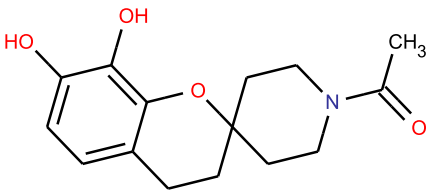
Data: ZINC20270898
Final_Score: 4.5

114



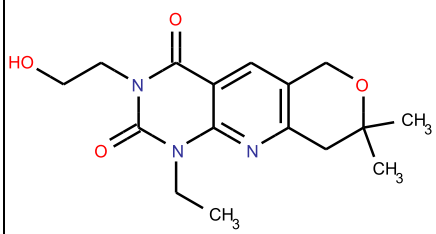
Data: ZINC19424226
Final_Score: 4.5

115



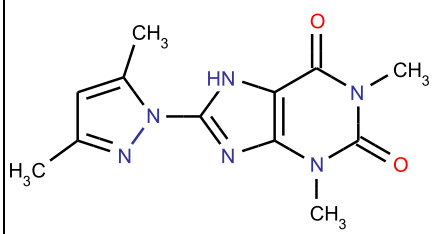
Data: ZINC08733263
Final_Score: 4.5

116



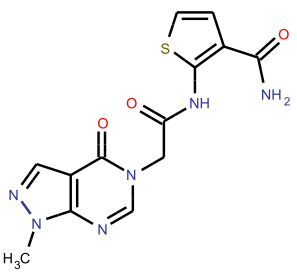
Data: ZINC27826339
Final_Score: 4.5

117



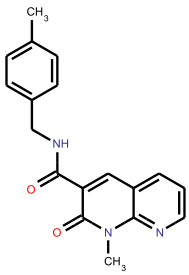
Data: ZINC00049541
Final_Score: 4.5

118



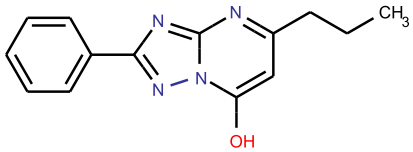
Data: ZINC24424955
Final_Score: 4.5

119

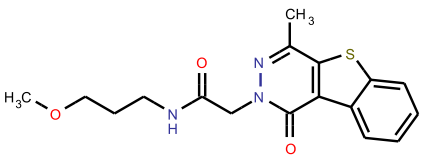
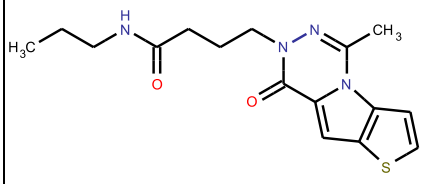
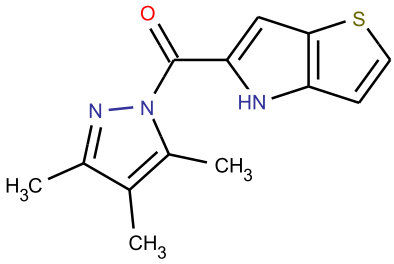
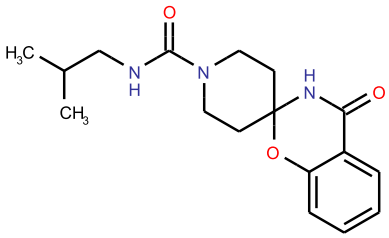
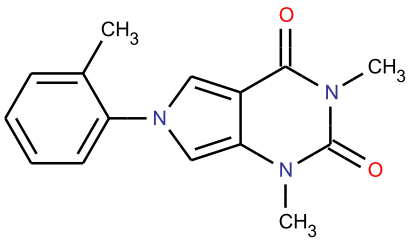
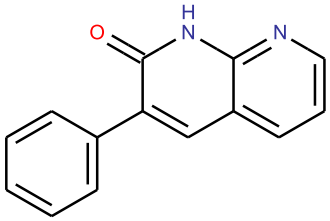
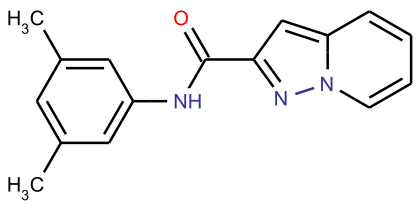
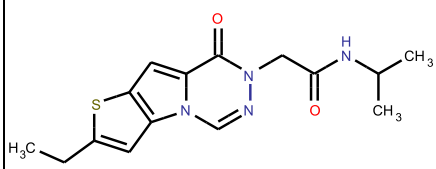
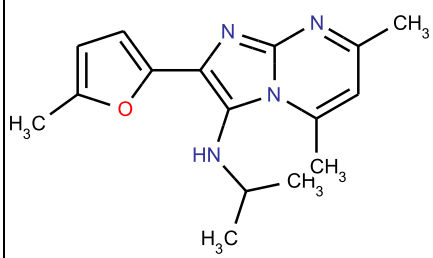
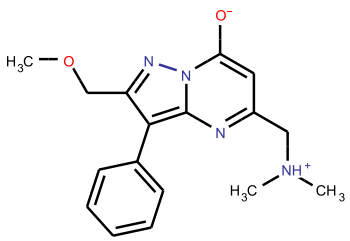
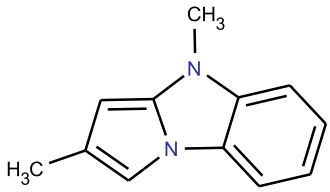
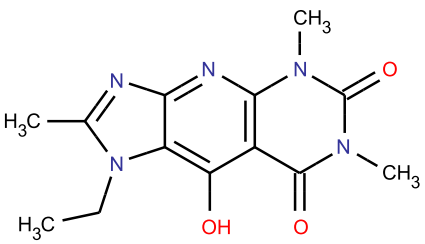
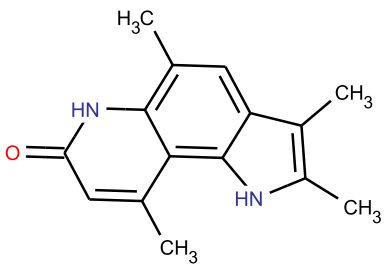
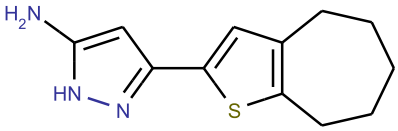
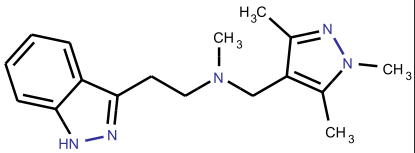


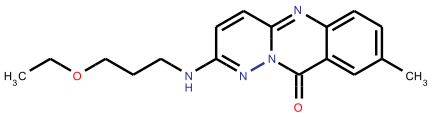
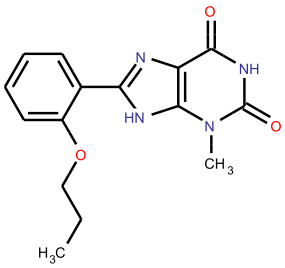
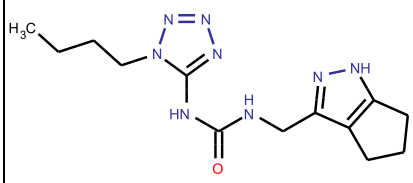
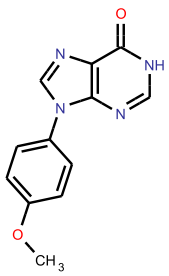
Data: ZINC05303547
Final_Score: 4.5

120



Data: ZINC16996405
Final_Score: 4.5

121	122	123
		
Data: ZINC06799016 Final_Score: 4.5	Data: ZINC05109147 Final_Score: 4.5	Data: ZINC06551306 Final_Score: 4.5
124	**125**	**126**
		
Data: ZINC21019152 Final_Score: 4.5	Data: ZINC00189001 Final_Score: 4.5	Data: ZINC19732006 Final_Score: 4.5
127	**128**	**129**
		
Data: ZINC05007347 Final_Score: 4.5	Data: ZINC06749536 Final_Score: 4.5	Data: ZINC04927560 Final_Score: 4.5
130	**131**	**132**
		
Data: ZINC28187792 Final_Score: 4.5	Data: ZINC26511999 Final_Score: 4.5 4.5	Data: ZINC08662726 Final_Score: 4.5
133	**134**	**135**
		
Data: ZINC04576464 Final_Score: 4.5	Data: ZINC19689510 Final_Score: 4.5	Data: ZINC19300429 Final_Score: 4.5

136	137	138
		
Data: ZINC05152290 Final_Score: 4.5	Data: ZINC06492257 Final_Score: 4.5	Data: ZINC23580883 Final_Score: 4.5
139		
		
Data: ZINC18217768 Final_Score: 4.5		